Universidade de Aveiro Departamento de Matemática
2015

Ana Márcia Correia Lapo

Modelação do preço de arrendamento dos imóveis destinados a habitação no concelho de Lisboa

Universidade de Aveiro Departamento de Matemática
2015

Ana Márcia Correia Lapo

Modelação do preço de arrendamento dos imóveis destinados a habitação no concelho de Lisboa

Relatório de estágio apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizado sob a orientação científica do Doutor Pedro Filipe Pessoa Macedo, Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro.

Dedico este trabalho à minha avó.

o júri

presidente

Prof. Doutor Agostinho Miguel Mendes Agra
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro.

Prof. Doutor António Carrizo Moreira
Professor Auxiliar do Departamento de Economia e Gestão Industrial da Universidade de Aveiro.

Prof. Doutor Pedro Filipe Pessoa Macedo
Professor Auxiliar do Departamento de Matemática da Universidade de Aveiro (orientador).

agradecimentos

Em primeiro lugar agradeço à minha família por ter sempre acreditado em mim.
Gostaria de agradecer ao meu orientador do Departamento de Matemática da Universidade de Aveiro, Professor Doutor Pedro Filipe Pessoa Macedo, pela disponibilidade e por todo o apoio prestados durante o período de realização deste trabalho.
Agradeço igualmente ao Engº Rui Costa, da empresa Ubiwhere onde decorreu o estágio, pela oportunidade e a todos os colaboradores pela recetividade e ajuda na integração.
Agradeço ainda à minha colega de estágio, Eva Karina Oliveira, pela amizade e ajuda.
Finalmente, agradeço ao Paulo de Carvalho por toda a paciência, compreensão e apoio.

palavras-chave Modelação, mercado imobiliário, regressão linear múltipla.

resumo

O objetivo deste trabalho é a construção de um modelo estatístico que permita identificar os atributos determinantes do preço de arrendamento dos imóveis situados no concelho de Lisboa. Para esse efeito são considerados quer atributos físicos típicos dos imóveis, como a área, a tipologia, o número de casas de banho, entre outros, quer atributos de localização. A informação recolhida respeitante aos últimos foi medida recorrendo aos tempos que se demora a percorrer as distâncias entre os imóveis e determinado ponto de interesse ou local de reconhecida influência, aquando da escolha de uma habitação para viver. A metodologia estatística selecionada foi a regressão linear múltipla. Mostraram-se relevantes para explicar o comportamento médio do preço de arrendamento dos imóveis algumas variáveis relacionadas com características intrínsecas e extrínsecas. No entanto, as últimas não tiveram o impacto esperado.

O modelo de regressão final obtido explica cerca de 65,6% da variação observada na variável dependente **Price**, em torno da sua média. Das onze variáveis nele incluídas, apenas três dizem respeito à localização geográfica do imóvel. Concluiu-se, igualmente, que os atributos físicos área útil, número de casas de banho e estado do imóvel são os que apresentam maior contribuição relativa para explicar o comportamento esperado do preço de arrendamento, dos imóveis destinados a habitação no concelho de Lisboa, conforme seria de esperar.

keywords Modelling, real estate market, multiple linear regression.

abstract

The purpose of this study is to build a statistical model that allows the identification of the main attributes influencing the rental price of a home, in the municipality of Lisbon. In order to do so, attributes related to the typical physical characteristics, as the area, typology, number of bathrooms, among others, and location attributes were taken into account. The information gathered in relation to the latter considered was the time required to cover the distance between a property and a point of interest or a place with recognized influence when individuals are looking for a place to live.

The statistical methodology applied was the multiple linear regression. The variables concerning both intrinsic and extrinsic characteristics have been proved relevant when it comes to explaining the average behavior of the rental price of properties. Nevertheless, extrinsic characteristics did not have the expected impact.

The final regression model obtained accounts for 65,6% of the variability observed in the outcome, the dependent variable **Price**.

Amongst the eleven variables it included, only three concern the location of the property. It was also concluded that the physical characteristics - as the useful area, number of bathrooms and the state of repair of the property - are the ones that have a greater impact to explain the expected behavior of the rental price, as expected.

Índice

Lista de Figuras

Lista de Quadros

Lista de Tabelas

1. Introdução

O presente trabalho consiste no relatório de estágio decorrido na empresa Ubiwhere durante um período de seis meses, de janeiro a junho de 2015. O estágio foi integrado no âmbito do projeto Livin'Lx, uma aplicação *web* que pretende sugerir ao utilizador que procura casa, o sítio ideal para viver em Lisboa, indo ao encontro do seu perfil e necessidades. O objetivo é a construção de um modelo estatístico que permita identificar os atributos determinantes do preço de arrendamento de um imóvel destinado a habitação, no concelho de Lisboa. A metodologia estatística escolhida para o efeito foi a regressão linear múltipla. Para alcançar o fim desejado foram consideradas não só características físicas dos imóveis, mas também a sua localização.

O processo exigiu a recolha de dados, com origem no portal imobiliário Imovirtual, que foram posteriormente tratados devidamente. Mostraram-se relevantes para explicar o comportamento médio do preço de arrendamento dos imóveis, variáveis relacionadas quer com os seus atributos físicos, quer com a sua localização. Verificou-se, no entanto, que as primeiras, atributos físicos, nomeadamente a área útil, o número de casas de banho e condição do imóvel, têm um maior impacto do que as segundas. Embora as características de localização se tenham mostrado importantes na formação dos preços, a sua contribuição não foi tão notória quanto se esperava.

Este relatório é constituído por seis capítulos. O presente capítulo inicia-se com a apresentação sucinta do problema e oferece um panorama geral das conclusões obtidas. Prossegue-se com uma breve apresentação da empresa. Posteriormente contextualiza-se o problema e define-se o objetivo do trabalho. No capítulo 2 expõe-se a teoria básica referente à metodologia estatística utilizada, a regressão linear múltipla. O capítulo 3 aborda a recolha e o tratamento dos dados disponibilizados pela empresa, assim como a sua análise descritiva. Nos capítulos 4 e 5 é aplicada a regressão linear múltipla para construir o modelo de regressão pretendido. Com o primeiro, capítulo 4, pretende fazer-se uma pesquisa exploratória inicial, para identificação das variáveis que devem ser incluídas no modelo final. O segundo, capítulo 5, trata esse modelo no que diz respeito à validação dos seus pressupostos, à análise dos coeficientes de regressão, entre outras questões. No capítulo 6 faz-se um breve resumo dos resultados obtidos com este estudo e apresentam-se sugestões para trabalho futuro.

1.1. A empresa

A empresa Ubiwhere teve início no ano de 2007 e tem a sua sede na cidade de Aveiro. Posteriormente abriu escritórios em São João da Madeira e, mais recentemente, em Coimbra. A sua localização estratégica, perto de duas reconhecidas universidades, tem vindo a permitir uma relação próxima com os centros académicos dando azo a projetos comuns e desafiantes.

Desde sempre a empresa apostou fortemente na inovação e na criatividade, na busca de novas e eficazes soluções na forma como interagimos com este mundo novo que nos rodeia, aproveitando a cada vez mais rápida evolução tecnológica.

Acompanhando a célere mudança de paradigma da realidade em que vivemos e a disponibilização de novas ferramentas, a Ubiwhere tem mantido o foco no futuro que, impreterivelmente, passa não só pela implementação de novas formas de comunicar, mas também pela renovação de comportamentos ou necessidades já existentes. Temos como exemplos das inúmeras áreas de ação onde esta jovem empresa singra, as telecomunicações, o turismo, a educação, os transportes, as energias e sustentabilidade, entre outros, utilizando as mais recentes tecnologias e conceitos para disponibilizar os melhores serviços e produtos. É nesta incessante procura de inovação e investigação que este estágio se desenrola.

Um ambiente descontraído e jovial promoveu uma rápida integração no seio da equipa e o espírito profissional proporcionou uma aprendizagem de eficiência nas metodologias de trabalho e dos fins a atingir. A sede onde decorreu este estágio localiza-se numa das mais prestigiadas zonas de Aveiro, com instalações e ambiência local agradáveis, contribuindo para um ainda maior empenho e produtividade dos seus colaboradores.

Foi no âmbito do projeto Livin'Lx que surgiu a parceria com a Universidade de Aveiro. A ideia subjacente a este serviço/produto, segundo os conceitos de *Big Data* e *Data Mining*, passa pela análise preditiva dos dados sobre a cidade de Lisboa e seus habitantes, direcionando o utilizador do serviço na procura das áreas/habitações que melhor se identificam com o seu perfil pessoal, social e profissional. Para este efeito a estatística tem-se revelado uma das melhores ferramentas na identificação dessas relações. Foi neste contexto que surgiu a oportunidade deste estágio. Seguindo os preceitos de qualidade da Universidade de Aveiro foi-me proposto alavancar o projeto Livin'Lx com o uso de metodologias estatísticas apreendidas ao longo do mestrado com o precioso suporte da referida instituição.

1.2. O problema

A opção de arrendar casa em detrimento da compra é nos dias de hoje uma realidade para muitas famílias. Os motivos que levam a esta situação podem ter origem diversa, nomeadamente profissional, financeira ou pessoal. Pode inclusivamente acontecer esta ser uma situação simplesmente mais confortável e lucrativa. De facto, a procura de uma habitação para viver torna-se uma tarefa muito menos complicada que a da compra. Basta atender ao nível de compromisso que se assume com um contrato de arrendamento de um imóvel quando comparado com a sua aquisição. Por outro lado, o arrendamento de uma habitação deixa para o senhorio uma série de formalidades ou encargos, responsabilidades com as quais o inquilino não tem de se inquietar. Falamos, por exemplo, da manutenção da casa e reuniões de condomínio. A despreocupação que o arrendamento proporciona ao inquilino permite igualmente que este se desloque facilmente consoante os seus interesses sociais e de lazer.

A situação profissional é, cada vez mais, um fator que motiva o arrendamento, pois os empregos são cada vez mais incertos, obrigando a deslocações frequentes na sua procura e consequente fixação habitacional. A capacidade económica da maioria das famílias não encaixa numa realidade de compra da habitação, caso surja a necessidade da sua alteração. O esforço financeiro exigido é, efetivamente, incomportável para a grande maioria das famílias.

1.2.1. Contextualização

O mercado habitacional distingue-se dos outros mercados de bens e serviços pela natureza do bem que é transacionado, a habitação. A sua importância é reconhecida, pois caracteriza-se por ser um bem de primeira necessidade, de custo elevado, pelo que, regra geral, a sua escolha exige uma reflexão muito ponderada por parte do comprador/arrendatário. Efetivamente, na nossa sociedade, a habitação constitui a grande fatia do património familiar e na qual é despendido uma parte considerável do orçamento mensal. A sua seleção assenta em duas perspetivas fundamentais e simultâneas, a saber: uma perspetiva da habitação, como bem de consumo, dado que se pretende tirar proveito direto de um local agradável para viver, e uma perspetiva da habitação como bem de investimento (1).

As especificidades do bem transacionado no mercado imobiliário, que o tornam tão particular, são essencialmente três: heterogeneidade, imobilidade e durabilidade (2). A heterogeneidade advém do facto de não haver dois imóveis exatamente iguais: as diferentes conjugações de diferentes atributos tornam uma habitação única. A imobilidade, ou seja, a

incapacidade de se poder deslocar uma habitação, será inevitavelmente um fator preponderante na tomada de decisão da sua escolha. Quando se opta por uma habitação não nos limitamos a adquirir um conjunto de atributos físicos, entendemos que a sua inserção numa vizinhança e o consequente usufruto de todo um conjunto de amenidades confere à habitação um caráter ímpar. Por outro lado, a associação obrigatória de uma localização a uma habitação dá origem ao assinalamento de submercados imobiliários, muitas vezes difíceis de delimitar (1). No que concerne à durabilidade, é um bem que pode durar várias décadas levando a que o *stock* de habitações novas seja reduzido quando comparado com o *stock* acumulado de habitações usadas. O impacto da durabilidade na dinâmica do mercado traduz-se, sem surpresa, num comportamento cíclico da procura, uma vez que qualquer decisão sobre bens desta natureza será facilmente adiada (3).

As especificações deste mercado acima expostas explicam a complexidade da atribuição de um valor a um imóvel habitacional, uma vez que este valor deve transparecer um conjunto de atributos intrínsecos e extrínsecos à habitação. A localização é, como já fizemos notar, um fator inseparável do produto imobiliário que se relaciona com o uso do solo na sua envolvente próxima, e assume um papel determinante nesta avaliação. Aquando da concretização desta avaliação, o avaliador deve reunir um vasto leque de informações sobre o imóvel, devendo ter em consideração diversos fatores, que podemos dividir em três categorias:

- Fatores gerais de localização- são dados de natureza internacional e nacional que podem influenciar o valor dos imóveis. A título de exemplo dos primeiros (natureza internacional) temos fatores tais como ameaças de guerra e a inflação mundial. Já de natureza nacional temos, por exemplo, as taxas de juro praticadas pelos bancos, rendimento *per capita*, crescimento anual da economia, entre outros. A um nível mais restrito podemos considerar dados relacionados com o município ou cidade em que a habitação está inserida. Desta última situação destacamos a dimensão e fase de crescimento da cidade, a tendência de crescimento da sua população e regulamentos municipais (4) .

- Fatores macro de localização- envolvem informação respeitante à densidade populacional, rede de transportes, distância ao centro, entre outros (3).

- Fatores micro de localização- afetam diretamente o valor da habitação. Referem-se a acessibilidades e facilidade de estacionamento, a existência e proximidade de transportes coletivos, de equipamentos, de serviços e espaços de lazer, tais como cafés, bares, farmácias, centros de saúde, hospitais, ATM's, correios, escolas primárias, infantários, escolas secundárias e universidades, esquadras de polícia, parques, locais de culto e igrejas, características da vizinhança

imediata (como a composição social do bairro ou existência de bairros sociais), características finais do produto (exposição solar e vistas), entre outros (4).

Não menos importantes, mas igualmente decisivos na determinação do preço do imóvel residencial, são os atributos específicos físicos e estruturais do imóvel. Incluímos neste grupo informação sobre a natureza da habitação (moradia ou apartamento), tipologia, áreas útil e total, número de casas de banho, existência de garagem, despensa, terraço, elevador, gás canalizado, alarme, estado de conservação, acabamentos, acessos a pessoas com mobilidade reduzida, condomínio fechado, entre outros.

1.2.2. Revisão bibliográfica

A procura da identificação dos atributos determinantes na formação do preço de uma habitação tem vindo a ser tema de estudo ao longo dos tempos, nomeadamente no ramo imobiliário. Este trabalho vai focar-se no preço de arrendamento de um imóvel, no concelho de Lisboa. A metodologia explorada será a dos modelos hedónicos, tendo por base a regressão linear múltipla. Os modelos de preços hedónicos tentam explicar a atribuição de determinado valor a um dado bem em função das suas características específicas. O primeiro estudo de modelos hedónicos desenvolvido na área da habitação é atribuído a Rosen (1974, citado por 1, p. 29). A relação próxima entre o valor de venda de um imóvel e o seu valor de arrendamento mensal é evidente, pelo que, a literatura consultada no contexto de modelos de preços hedónicos, maioritariamente dirigida ao preço de venda de um imóvel, ofereceu uma orientação indiscutível ao estudo efetuado.

No âmbito do mercado imobiliário, as variáveis explicativas do preço presentes neste tipo de modelos incluem, não só, mas também, essencialmente duas vertentes: uma que engloba as características físicas do imóvel e outra que engloba a sua localização. Parece ser consensual que, incorporar aspetos espaciais, ou seja, medir a influência da localização é, sem dúvida, a parte mais difícil de concretizar.

São inúmeros os trabalhos de investigação que se têm debruçado sobre esta problemática. Por exemplo, o estudo desenvolvido por Tarré (5) visou analisar se duas zonas com particularidades distintas da Cidade de Lisboa têm valores de avaliação por metro quadrado diferenciados, no âmbito do crédito à habitação. Também neste estudo, pesquisou-se se as variáveis com capacidade explicativa envolvidas nos dois modelos hedónicos obtidos, um para cada zona, seriam iguais e qual a sua contribuição relativa para explicar a variação do preço observado em cada uma delas. As variáveis consideradas neste trabalho incluíram, para além de uma descrição física típica da habitação, uma descrição da sua envolvente próxima. A autora conclui que, efetivamente, a

localização é fator preponderante na valorização do bem imóvel, uma vez que os resultados apontaram no sentido de haver diferenças nos preços de avaliação por metro quadrado nas duas zonas. Já Batista (1) procurando também a identificação dos fatores determinantes do preço da habitação, efetua duas análises, uma numa escala macro e outra numa escala micro. A primeira incluiu indicadores das características das habitações, ao nível municipal, localizadas nos municípios de Portugal continental. As variáveis recolhidas referem-se a aspetos que caracterizam genericamente a vizinhança das habitações e a aspetos estruturais dominantes. A análise micro incidiu sobre os municípios de Ílhavo e Aveiro. A informação reunida compreendeu atributos físicos básicos (área, preço, tipologia, etc.), atributos físicos descritivos (varanda, garagem, duplex, etc.) e atributos espaciais. Neste estudo o autor mede as variáveis de localização à custa do conceito de potencial. Este engloba a quantidade de pontos de interesse e a distância mínima desde um centróide de uma microzona (também previamente determinados) até ao ponto de interesse desejado. Nesta criação de variáveis de localização é ainda explorado o nível de importância destes pontos de interesse. Distribuídos por categorias, eles são considerados mais ou menos relevantes se são contabilizados dentro de um raio de 600 metros, 1200 metros ou 1800 metros. O estudo concluiu que, a par dos atributos físicos, a localização é preponderante na determinação do preço dos imóveis.

Na tentativa de explicar a formação do valor de arrendamento de imóveis na cidade de Porto Alegre, Gonzalez e Formoso (6) recolhem informação respeitante a atributos intrínsecos típicos dos imóveis (área, tipologia, estado, etc.) e a atributos extrínsecos de localização, nomeadamente distâncias a supermercado, ao centro, a centros comerciais, proximidade de favelas, entre outros. Novamente, o modelo final obtido conclui que, tal como as esperadas particularidades físicas de um imóvel, também a sua localização tem influência na formação do preço de arrendamento.

Similarmente, o trabalho desenvolvido por Guedes (7) propôs-se a identificar os atributos que têm impacto na formação do preço por metro quadrado de um imóvel. O estudo propõe oito combinações de variáveis explicativas, portanto oito modelos, e conclui que o que apresenta melhor desempenho inclui a variável de localização considerada.

Refere-se ainda um trabalho digno de interesse desenvolvido no concelho de Gaia (8). Neste estudo o concelho é dividido em quatro zonas e para cada uma delas são incorporadas variáveis explicativas que caracterizam a localização, a dimensão e tipo de habitação, a qualidade de acabamentos e equipamentos, o equilíbrio das partes constituintes do imóvel e a comercialização. As variáveis respeitantes à localização são específicas de cada zona considerada. Dos resultados

destaca-se que, em todas as zonas, os aspetos referentes à localização influenciam o preço por metro quadrado, não acontecendo o mesmo com os aspetos de comercialização.

A investigação levada a cabo por Tavares, Moreira e Pereira (9) inclui, para além de outras variáveis independentes, variáveis relacionadas com a localização e com as vistas de apartamentos situados em dois empreendimentos, um turístico, em Tróia, e outro residencial, situado em Espinho, numa tentativa de tentar perceber o efeito que causam no preço por metro quadrado. As conclusões obtidas em ambos os casos dizem que as vistas são efetivamente um fator influente na formação do indicador de preço considerado.

Para finalizar, ainda relacionado com esta temática da localização, sugere-se a consulta de um artigo dos mesmos autores, onde se faz uma revisão bibliográfica sobre um vasto conjunto de estudos que pesquisaram em que medida as características da envolvente próxima de um imóvel podem influenciar a sua avaliação imobiliária, as denominadas externalidades (10). Neste artigo são apresentados três quadros resumo compilando a seguinte informação: autores que realizaram estudos sobre externalidades positivas e negativas que têm um impacto positivo e negativo, respetivamente, na avaliação da habitação, e um quadro resumo das externalidades que são consideradas simultaneamente positivas e negativas e respetivos autores.

1.2.3. Objetivo

O objetivo deste trabalho é a construção de um modelo estatístico que permita identificar os atributos que influenciam o valor de um imóvel no mercado de arrendamento, no concelho de Lisboa, servindo de base a uma ferramenta de *profiling* e avaliação de imóveis. Serão considerados fatores intrínsecos e extrínsecos ao imóvel, referindo-se os primeiros a características físicas e estruturais e os segundos a especificidades de localização.

2. Modelo de regressão linear múltipla

O modelo de regressão linear múltipla procura predizer o comportamento de uma variável dependente (ou resposta), y, a partir de um conjunto de variáveis independentes (ou explicativas), $x_2, x_3, \dots, x_k$, assumindo uma relação do tipo

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u\,, \qquad [1]$$

onde:

- a variável aleatória y se designa por regressando;
- a variável aleatória x_j $(j = 2, 3, \dots, k)$ é um regressor;
- o parâmetro β_j $(j = 1, 2, \dots, k)$ chama-se coeficiente de regressão;
- u é a variável residual, não observável, que abrange todos os fatores que não são considerados no modelo, mas que podem afetar o comportamento da variável explicada.

Na presença de dados seccionais (observações de certos atributos de certas entidades em determinado momento), como é o caso, uma amostra de dimensão n da população é constituída por n entidades observáveis para cada variável presente no modelo (11). Considerando uma amostra de dimensão n, $\{(y_t, x_{t2}, x_{t3}, \dots, x_{tk}) : t = 1, 2\ \dots, n\}$, definem-se n relações amostrais a partir do modelo [1]:

$$y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk} + u_t\,,\ (t = 1, 2, \dots, n)\,, \qquad [2]$$

onde u_t é a variável aleatória residual (11).

Em alternativa, as n igualdades representadas na equação [2] podem ser apresentadas usando a notação matricial,

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ X = \begin{bmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{nk} \end{bmatrix},\ \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix},\ U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}.$$

Obtém-se, assim, a relação amostral $Y = X\beta + U$, onde:

- Y é o vetor $n \times 1$ das observações (aleatórias) do regressando;

- X é a matriz $n \times k$ das observações (aleatórias) dos regressores;

- β é o vetor $k \times 1$ dos coeficientes de regressão;

- U é o vetor $n \times 1$ das variáveis residuais.

2.1. Hipóteses básicas do modelo de regressão linear

Apresentam-se a seguir os pressupostos de aplicação do modelo de regressão linear (11):

- o valor esperado da variável residual, u_t, condicionado pela matriz dos regressores, X, é nulo, ou seja, $E(u_t|X) = 0 \ \ (t = 1, 2, \dots, \ n)$. Esta hipótese permite concluir que o valor esperado de y_t é dado por $E(y_t|X) = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \cdots + \beta_k x_{tk}$;

- a variância da variável residual, u_t, condicionada pelas observações dos regressores é constante, ou seja, tem-se $Var(u_t|X) = \sigma^2 > 0 \ \ (t = 1, 2, \dots, \ n)$;

- ausência de autocorrelação, isto é, $Cov(u_t, u_s|X) = 0 \ \ (t, s = 1, 2, \dots, \ n; \ t \neq s)$;

- a característica da matriz X é igual a k e $k < n$: $r(X) = k < n$. Com esta hipótese as colunas da matriz X são linearmente independentes, ou seja, o vetor das observações de um regressor não é combinação linear dos vetores das observações de outros regressores.

O modelo de regressão que verifique os pressupostos mencionados designa-se por modelo de regressão linear clássico com parâmetros desconhecidos β_1, β_2, …, β_k e σ^2.

2.2. Estimação dos coeficientes de regressão pelo método dos mínimos quadrados

Para estimar os coeficientes de regressão, $\beta_j, j = 1, 2, \dots, k$, recorre-se usualmente ao método dos mínimos quadrados. Este método consiste em minimizar a soma dos quadrados dos resíduos, onde um resíduo relativamente à observação t se define por $\tilde{u}_t = y_t - \left(\tilde{\beta}_1 + \tilde{\beta}_2 x_{t2} + \tilde{\beta}_3 x_{t3} + \cdots + \tilde{\beta}_k x_{tk}\right)$, sendo $\tilde{\beta}$ um valor qualquer de β (11). O estimador dos mínimos quadrados de β, b, é aquele que minimiza a função $\varphi\left(\tilde{\beta}\right) = \sum_{t=1}^{n} {\tilde{u}_t}^2$. Prova-se que $b = \ (X^T X)^{-1} X^T Y$.

Depois de estimados os coeficientes de regressão, obtém-se a equação de regressão linear ajustada,

$$\hat{y}_t = b_1 + b_2 x_{t2} + b_3 x_{t3} + \cdots + b_k x_{tk} \ \ (t = 1, 2, \ldots , \ n).$$

2.3. Coeficiente de determinação

Para avaliar a qualidade do ajustamento dos valores preditos pelo modelo, $\hat{y}_t$, aos dados, y_t, $(t = 1, 2, \ldots , \ n)$, usa-se como indicador o coeficiente de determinação, R^2. Este coeficiente mede a proporção da variabilidade total que é explicada pelo modelo de regressão e é definido por

$$R^2 = \frac{VE}{VT} = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2},$$

onde $\bar{y}$ é a média dos y_t.

Tem-se que $0 \leq R^2 \leq 1$, sendo que quanto mais próximo de 1 melhor é o grau de ajustamento do modelo. Contudo, este indicador tem o inconveniente de nunca decrescer sempre que se acrescenta um regressor, mesmo que tenha uma influência reduzida sobre a variável dependente (11). Para contornar esta desvantagem recorre-se ao coeficiente de determinação ajustado, $\overline{R^2}$, assim definido

$$\overline{R^2} = 1 - \frac{VR/(n-k)}{VT/(n-1)} = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2/(n-k)}{\sum_{t=1}^{n}(y_t - \bar{y})^2/(n-1)}.$$

O coeficiente de determinação ajustado aumenta apenas se o acréscimo de um regressor ao modelo conduzir a uma maior proximidade entre as observações da variável dependente e os respetivos valores ajustados (11).

2.4. Inferência estatística

Com o objetivo de fazer procedimentos de inferência estatística sobre os parâmetros do modelo de regressão linear, nomeadamente a realização de testes de hipóteses e a construção de intervalos de confiança, vai admitir-se que a variável residual condicionada por X segue uma distribuição Normal com valor esperado 0 e variância constante, isto é, $u_t|X \sim N(0, \sigma^2)$ (11).

No contexto da análise do modelo de regressão linear obtido, em particular a realização de inferência estatística, destacam-se dois testes de hipóteses que se apresentam a seguir. O primeiro, designado por teste de significância global da regressão, pretende averiguar se o modelo ajustado é ou não significativo, ou, dito de outra forma, se o modelo proposto é adequado para explicar o comportamento da variável dependente (11). Estão em causa as seguintes hipóteses:

$$H_0: \beta_2 = \beta_3 = \cdots = \ \beta_k = 0$$

$$vs.$$

$$H_1: \exists\, \beta_j \neq 0\ (j = 2,3, \dots, k).$$

A estatística de teste é dada por

$$F = \frac{VE/(k-1)}{VR/(n-k)} = \frac{\sum_{t=1}^{n}(\hat{y}_t - \bar{y})^2/(k-1)}{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2/(n-k)} \sim F(k-1, n-k)\,.$$

Rejeita-se a hipótese nula se o $p - value = P(F > F_{obs}|H_0)$ associado for inferior ou igual a α, o nível de significância do teste (11).

Finalmente, para fazer inferência sobre um coeficiente de regressão isolado recorremos ao teste seguinte:

$$H_0: \beta_j = \beta_j^0$$

$$vs.$$

$$H_1: \beta_j \neq \beta_j^0\ (j = 1,2,3, \dots, k).$$

A estatística de teste envolvida é dada por

$$t_j = \frac{b_j - \beta_j^0}{s_{b_j}} \sim t(n-k)\,,$$

onde s_{b_j} é o erro padrão de b_j, e $s_{b_j} = \sqrt{s^2 m^{jj}}$, sendo m^{jj} o elemento diagonal de ordem j da matriz $(X^T X)^{-1}$ e $s^2 = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n-k}$. Dá-se particular interesse ao caso em que $\beta_j^0 = 0$. Se se rejeitar a hipótese nula diz-se que o coeficiente β_j é estatisticamente significativo e a variável independente associada é relevante para explicar o comportamento médio de y_t (11).

2.5. Métodos de seleção de variáveis

A escolha do conjunto de regressores mais importantes a incluir no modelo pode ser efetuada por diferentes métodos de seleção. Apresentam-se de seguida os três procedimentos disponibilizados no SPSS Statistics:

- seleção *backward*: neste método o modelo inicial inclui todos os regressores. Para cada um deles é calculado o valor da estatística F. A variável que apresentar o menor valor de F é removida do modelo se esse valor for inferior a outro estabelecido previamente,

$F_{Removal}$. No passo seguinte repete-se o processo, considerando agora o modelo ajustado com menos um regressor. O procedimento termina quando todas as variáveis presentes no modelo possuem um valor da estatística F superior ao valor crítico $F_{Removal}$ (12).

- seleção *forward*: neste método a primeira variável incluída no modelo é a que apresenta o maior coeficiente de correlação (em valor absoluto) com a variável dependente. As próximas variáveis a ser introduzidas no modelo, de forma sequencial, são aquelas que tiverem maior coeficiente de correlação parcial entre a variável dependente e aquela que se pretende incluir, considerando os efeitos das variáveis anteriormente inseridas. O coeficiente de correlação parcial referido mede, portanto, a relação entre y e x_j depois de eliminar os efeitos provocados pelas variáveis já introduzidas. Admitindo que $x_{t_1}, \dots, x_{t_l}$ já são variáveis envolvidas no modelo, o coeficiente de correlação parcial é o coeficiente de correlação entre dois tipos de resíduos: os resíduos resultantes da regressão de y sobre $x_{t_1}, \dots, x_{t_l}$, e os resíduos obtidos da regressão de x_j sobre $x_{t_1}, \dots, x_{t_l}$ (13). Para cada nova variável que é testada é calculado o valor da estatística F. Se esse valor for superior a um valor crítico de entrada, F_{Entry}, então a variável é adicionada ao modelo e é eliminada caso contrário (12).

- seleção *stepwise*: este método mistura os dois métodos anteriormente expostos. Começa como o método *forward*, no entanto, após a introdução de uma nova variável é feita uma reavaliação às variáveis inseridas até então, usando o método *backward*. Deste modo, é possível remover uma variável que vê a sua importância reduzida pela adição de novas variáveis. O procedimento termina quando nenhuma variável independente consegue entrar no modelo, com base no valor de F_{Entry} e nenhuma é eliminada com base no valor de $F_{Removal}$ (12).

2.6. Previsão

Um dos principais fins da obtenção de um modelo de regressão linear é o da previsão de novos valores do regressando a partir de concretizações dos valores dos regressores. A designada previsão em média corresponde à estimação do valor esperado do regressando para um determinado vetor de observações das variáveis independentes (11). Ou seja, considerando que se fixam valores para os regressores, $c = [1 \quad c_2 \quad c_3 \quad \dots \quad c_k]$, pretende-se estimar o parâmetro

$$\theta = E(y_t | \boldsymbol{x}_{t2} = c_2, \boldsymbol{x}_{t3} = c_3, \dots, \boldsymbol{x}_{tk} = c_k) = \beta_1 + \beta_2 c_2 + \beta_3 c_3 + \cdots + \beta_k c_k \, .$$

Um intervalo de confiança de grau $1-\alpha$ para o valor médio, θ, é dado por

$$\left]\hat{\theta}-t_{\frac{\alpha}{2}}\,s\sqrt{c(X^TX)^{-1}c^T};\hat{\theta}+t_{\frac{\alpha}{2}}\,s\sqrt{c(X^TX)^{-1}c^T}\right[,$$

onde $\hat{\theta}=b_1+b_2c_2+b_3c_3+\cdots+b_kc_k$ (11).

Faz-se previsão pontual quando se pretende prever um valor particular do regressando a partir de um conjunto de observações das variáveis independentes. Deste modo, para concretizações dos valores das variáveis independentes, pode obter-se uma estimativa pontual de y_t substituindo os valores dos regressores na equação do modelo ajustado previamente aos dados. Considere-se novamente $c=[1 \quad c_2 \quad c_3 \quad \cdots \quad c_k]$ e ainda $y_0=\beta_1+\beta_2c_2+\beta_3c_3+\cdots+\beta_kc_k+u_0$, onde u_0 é a variável residual correspondente. O respetivo modelo estimado (mínimos quadrados) é dado por $\hat{y}_0=b_1+b_2c_2+b_3c_3+\cdots+b_kc_k$ (11). Um intervalo de confiança de grau $1-\alpha$ para y_0 é dado por

$$\left]\hat{y}_0-t_{\frac{\alpha}{2}}\,s\sqrt{1+c(X^TX)^{-1}c^T};\hat{y}_0+t_{\frac{\alpha}{2}}\,s\sqrt{1+c(X^TX)^{-1}c^T}\right[.$$

3. Base de dados

No presente capítulo pretende-se expor todo o trabalho realizado para a obtenção da base de dados que foi alvo de estudo neste relatório. Durante esta fase recorreu-se ao uso do *software* SPSS Statistics, versão 22. Numa fase inicial de recolha de dados, quando houve necessidade de agregar informação contida em bases diferentes, fez-se também uso da ferramenta *Excel* do *Microsoft Office*. Inicialmente apresenta-se a forma como se tratou a informação recolhida. Posteriormente apontam-se algumas dificuldades que foram surgindo ao longo do processo e, para finalizar, faz-se a sua análise descritiva.

3.1. Recolha e tratamento

Os dados recolhidos e fornecidos pela empresa têm origem no portal imobiliário Imovirtual. Os 7852 registos constantes da base de dados inicial, obtida em março do corrente ano, referem-se a imóveis situados exclusivamente no concelho de Lisboa. Tendo em conta o objetivo definido pela empresa, ou seja, identificar os atributos determinantes do preço de um imóvel destinado a habitação, foram selecionados apenas os destinados a venda e arrendamento, do tipo moradia e apartamento. O passo seguinte consistiu na remoção de casos duplicados. A base de dados inicial foi então dividida em duas: uma base para venda e outra para arrendamento. Neste relatório será tratada a última, ou seja, imóveis respeitantes ao negócio de arrendamento. Obteve-se assim uma base de dados com 1278 imóveis.

A introdução de um imóvel no portal pode ser efetuada por diversos agentes, nomeadamente profissionais do ramo do mercado imobiliário e particulares. Durante este processo é pedido ao utilizador que preencha um formulário *online*, onde constam uma série de campos, uns obrigatórios, outros não, que têm como finalidade a descrição do imóvel. A informação reunida refere-se essencialmente aos seus atributos físicos e à sua localização.

Dos campos de preenchimento obrigatório fazem parte os atributos preço, natureza, tipo de negócio, localização, tipologia e área útil. No que diz respeito à localização, a informação mínima exigida requer o código postal do imóvel. Para cada imóvel, os valores atribuídos às variáveis **Price**, **HouseType**, **NrRooms** e **UsefulArea** descrevem-no relativamente ao preço mensal de arrendamento, natureza, tipologia e área útil (em metros quadrados), respetivamente. A restante

informação recolhida, sem carácter obrigatório, é descrita na Tabela 1 apresentando-se paralelamente as variáveis relativas a cada atributo e que serão usadas doravante.

Tabela 1- Variáveis (não obrigatórias) e respetiva descrição.

Variável	*Descrição*
TotalArea	Área bruta, em metros quadrados
State	Condição do imóvel
Energy	Certificado energético
Year	Ano de construção
NrWCs	Número de casas de banho

Durante a inserção do imóvel no portal é dada a possibilidade do utilizador selecionar certas características deste imóvel, de entre um conjunto de vinte e três: acessibilidades a pessoas com mobilidade condicionada, alarme, aquecimento central, ar condicionado, árvores de fruto, condomínio fechado, cozinha equipada, elevador, estacionamento, garagem, hidromassagem/jacuzzi, jardim, lareira, mobilado, piscina, quintal/horta, imóvel de banco, som ambiente, varanda, vigilância/segurança, vista de campo/serra, vista de mar e vista de rio. Assumindo que estas características acrescentam valor ao imóvel, criou-se uma nova variável, **soma_caracteristicas**, que, tal como o nome sugere, atribui a cada imóvel um número entre 0 e 23. Esta variável pretende estabelecer uma distinção entre os imóveis, ficando melhor classificado aquele que reunir mais características. Pelo exposto, entende-se que a não seleção destes atributos implica a sua inexistência, só havendo vantagens em mencioná-los.

Criou-se ainda uma variável quantitativa, **Age**, obtida calculando a diferença entre o ano de 2015 e o ano de construção de cada imóvel. Esta variável fornece, naturalmente, a idade do imóvel.

Por fim esclarece-se que no portal *online* Imovirtual, o campo "Condição" permite a seleção de um elemento do conjunto {"Novo", "Renovado", "Usado", "Em construção", "Para recuperar", "Em ruína"}. Os registos constantes da base de arrendamento abarcam apenas os casos "Novo", "Renovado", "Usado" e "Em ruína". Procedeu-se então a uma redefinição da condição de cada imóvel, considerando-se a já referida variável **State**, a qual assume somente dois tipos de condição: "Novo" e "Usado". Portanto, todos os registos com "Renovado", "Usado" e "Em ruína" são considerados como "Usado".

No que concerne aos atributos de localização da habitação, após uma pesquisa inicial focada essencialmente em artigos sobre modelação de preços no mercado imobiliário, foram selecionados um conjunto de pontos de interesse cuja proximidade ao imóvel podem afetar o seu preço de arrendamento. Os pontos de interesse identificam elementos de benefício geral, como equipamentos, comércio, parques, serviços públicos, locais de culto, entre outros. Tendo em conta a influência clara da existência de determinados locais como zonas industriais, aterros e refinarias químicas nas imediações de um imóvel, estes foram igualmente considerados neste trabalho. Numa tentativa de incluir esta informação no estudo, e desta forma descrever a vizinhança de cada imóvel, procedeu-se à criação de novas variáveis relativas a distâncias e tempos. Após uma exposição à empresa, pesando as suas limitações e gestão de recursos, foi possível, para cada registo, através do seu código postal, obter a distância, em quilómetros, entre o imóvel e o lugar de referência desejado mais próximo, assim como o respetivo tempo, em minutos, que se demora a percorre-la de carro. A Tabela 2 compila a informação acabada de expor e, à semelhança da anterior, apresenta as respetivas variáveis:

Tabela 2- Variáveis respeitantes a tempos e distâncias e respetiva descrição.

Variável	*Descrição*
Time_bus_station	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e a estação de comboios mais próxima.
Time_subway	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e a estação de metro mais próxima.
Time_cafe	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e o café mais próximo.
Time_grocery	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e a mercearia mais próxima.
Time_gym	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e o ginásio mais próximo.
Time_hospital	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e o hospital mais próximo.
Time_night_club	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e o estabelecimento de diversão noturna mais próximo.
Time_park	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e o parque mais próximo.
Time_pharmacy	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e a farmácia mais próxima.
Time_restaurant	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e o restaurante mais próximo.
Time_school	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e o estabelecimento de ensino mais próximo.
Time_aterro	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e o aterro mais próximo.
Time_ZI	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e a zona industrial mais próxima.
Time_refinaria	Tempo, em minutos, que demora a percorrer de carro, a distância entre o imóvel e a refinaria situada em Algés.
dist_bus_station	Distância, em quilómetros, entre o imóvel e a estação de comboios mais próxima.

dist_subway	Distância, em quilómetros, entre o imóvel e a estação de metro mais próxima.
dist_cafe	Distância, em quilómetros, entre o imóvel e o café mais próximo.
dist_grocery	Distância, em quilómetros, entre o imóvel e a mercearia mais próxima.
dist_gym	Distância, em quilómetros, entre o imóvel e o ginásio mais próximo.
dist_hospital	Distância, em quilómetros, entre o imóvel e o hospital mais próximo.
dist_night_club	Distância, em quilómetros, entre o imóvel e o estabelecimento de diversão noturna mais próximo.
dist_park	Distância, em quilómetros, entre o imóvel e o parque mais próximo.
dist_pharmacy	Distância, em quilómetros, entre o imóvel e a farmácia mais próxima.
dist_restaurant	Distância, em quilómetros, entre o imóvel e o restaurante mais próximo.
dist_school	Distância, em quilómetros, entre o imóvel e o estabelecimento de ensino mais próximo.
dist_aterro	Distância, em quilómetros, entre o imóvel e o aterro mais próximo.
dist_ZI	Distância, em quilómetros, entre o imóvel e a zona industrial mais próxima.
dist_refinaria	Distância, em quilómetros, entre o imóvel e a refinaria situada em Algés.

Todas as variáveis respeitantes a distâncias apresentam um elevado número de zeros, significando o valor "0" que a distância é menor do que $1\ Km$. Tal facto levou a que muitos imóveis ficassem igualmente classificados, não havendo, portanto, grande variabilidade nestas variáveis. Esta situação não traz qualquer valor ao estudo. Por este motivo optou-se por usar a informação contida nas distâncias do seguinte modo: definiu-se uma nova variável quantitativa, **NrInteresse**, que contabiliza os serviços e amenidades existentes na vizinhança de um imóvel, num raio de $1\ Km$. Para este efeito não contribuíram somente as variáveis **dist_aterro**, **dist_ZI** e **dist_refinaria**. Convém ressalvar que esta contabilização foi efetuada atendendo às anotações de pontos de interesse já constantes do portal que está a ser desenvolvido pela empresa.

Depois de organizada toda a informação respeitante a cada imóvel de forma clara e fiável, procedeu-se a uma análise geral e informal dos dados por parte dos elementos do grupo de trabalho. Por ser demasiado evidente a presença de um elevado número de imóveis com preço mensal de arrendamento demasiado reduzido, ficou estabelecido que apenas seriam considerados os imóveis cujos valores da variável **Price** fossem maiores ou iguais a 100 Euros. Foram igualmente removidos alguns casos para os quais se verificou incoerência de informação.

Pelo exposto, a base final alvo deste trabalho de investigação consiste num total de 760 imóveis. Para cada um deles foram observadas vinte e cinco variáveis. Destas, vinte e duas são quantitativas: **Price**, **NrRooms**, **UsefulArea**, **TotalArea**, **Age**, **NrWCs**, **soma_caracteristicas**,

Time_bus_station, **Time_subway**, **Time_cafe**, **Time_grocery**, **Time_gym**, **Time_hospital**, **Time_night_club**, **Time_park**, **Time_pharmacy**, **Time_restaurant**, **Time_school**, **Time_aterro**, **Time_ZI**, **Time_refinaria**, **NrInteresse**. As restantes são qualitativas, sendo **State** e **HouseType** nominais e **Energy** ordinal. As codificações adotadas para as variáveis qualitativas encontram-se na Tabela 3.

Tabela 3- Codificações adotadas para as variáveis qualitativas.

Variável	*Código*	*Legenda*
State	0	Usado
	1	Novo
HouseType	0	Moradia
	1	Apartamento
Energy	1	A+
	2	A
	3	B
	4	B-
	5	C
	6	D
	7	E
	8	F
	9	G

Como principal dificuldade no processo de recolha dos dados assinala-se a presença de valores omissos, por vezes em elevado número, na maioria das variáveis. O motivo explica-se pela inexistência de uniformização da informação que é inserida para cada imóvel.

3.2. Fragilidades dos dados

No decurso do processo da recolha de informação sobre os imóveis, fase da exclusiva responsabilidade da empresa, foram surgindo algumas dificuldades, limitações e obstáculos que se tentaram contornar da melhor forma possível. Assinalam-se os seguintes:

- falta de uniformização no preenchimento dos campos descritivos da habitação no formulário *online*. A não obrigatoriedade do preenchimento de todos os campos por parte do utilizador levou à presença de valores omissos na maioria das variáveis, sendo em elevado número em algumas delas;
- foram encontradas algumas incoerências entre o texto livre incluído no campo relativo à descrição do imóvel e os campos correspondentes às variáveis consideradas no estudo;
- a falta de conhecimento da zona impossibilitou a determinação exaustiva de todos os pontos de interesse, refinarias, zonas industriais e aterros do concelho de Lisboa;

- uma vez que a finalidade da publicação do imóvel é o arrendamento, o utilizador pode ocultar ou alterar atributos que promovam a sua desvalorização;
- existência de imóveis iguais. Tal situação pode efetivamente corresponder à realidade ou dever-se à não exclusividade da publicação. Ou seja, o mesmo imóvel pode estar a ser publicado por agências imobiliárias diferentes ou pelo mesmo proprietário, com nomes de utilizador diferentes;
- imóveis muito semelhantes podem apresentar valores de arrendamento muitos díspares, de acordo com o agente que está a publicitá-los;
- fatores pessoais, tais como necessidades financeiras, podem conduzir a que os apartamentos sejam subvalorizados ou, pelo contrário, sobrevalorizados;
- em muitos casos os preços observados podem não corresponder ao real valor de transação;
- falta de garantia de representatividade da amostra.

Prossegue-se na secção seguinte à análise preliminar dos dados.

3.3. Análise descritiva

Nesta fase pretende-se fazer a descrição das variáveis presentes na amostra recolhida e sintetizar a estrutura da sua distribuição. A exploração realizada poderá levar à deteção de valores estranhos que serão individual e pormenorizadamente avaliados. Para este efeito recorre-se à análise de estatísticas e a representações gráficas como histogramas, diagramas de extremos e quartis, diagramas de dispersão, entre outros. Inicia-se esta análise preliminar dos dados pelas variáveis qualitativas: **HouseType**, **State** e **Energy**.

A base de dados recolhida contém 760 imóveis, distribuídos por 24 freguesias de Lisboa, conforme o diagrama de barras da Figura 1. Constata-se que a maioria dos imóveis está localizado na freguesia de Arroios.

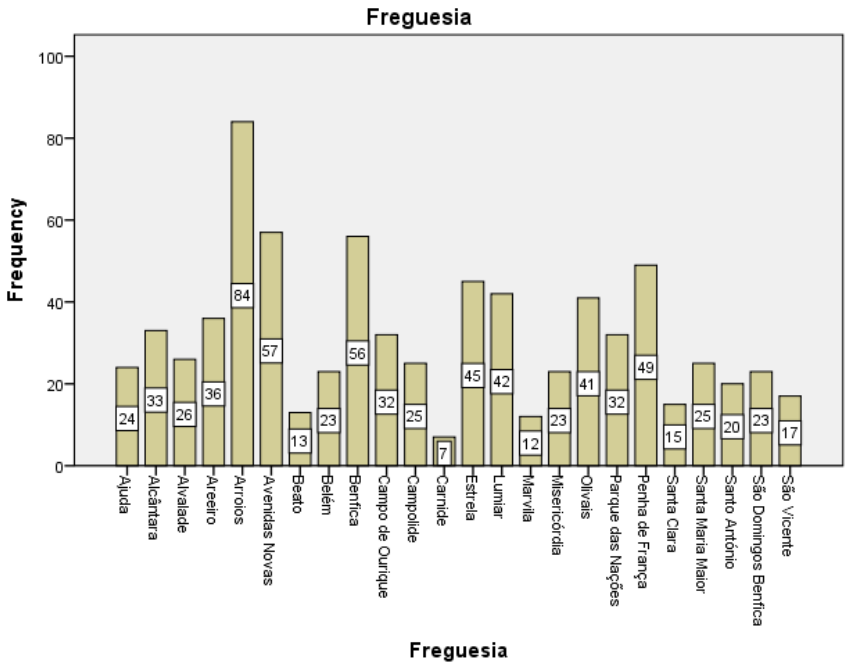

Figura 1- Distribuição geográfica dos imóveis no concelho de Lisboa.

Destes 760 imóveis, 12 são do tipo moradia e 748 do tipo apartamento (ver Quadro 1). As análises efetuadas neste estudo que envolvem a variável **HouseType** poderão ser algo frágeis pois constata-se que, nesta amostra, o número de apartamentos é consideravelmente superior ao número de moradias. 89.9% representam imóveis usados e 5% são novos; para os restantes 5.1% dos imóveis, correspondendo a 39 registos, não há conhecimento sobre a sua condição, ou seja, a variável **State** apresenta 39 valores omissos (ver Quadro 2).

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Moradia	12	1,6	1,6	1,6
	Apartamento	748	98,4	98,4	100,0
	Total	760	100,0	100,0	

*Quadro 1- Frequências da variável **HouseType**.*

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Usado	683	89,9	94,7	94,7
	Novo	38	5,0	5,3	100,0
	Total	721	94,9	100,0	
Missing	-1	39	5,1		
Total		760	100,0		

*Quadro 2- Frequências da variável **State**.*

A variável **Energy** é das variáveis que apresenta maior número de valores omissos; mais precisamente, para 471 imóveis não foi possível obter informação sobre a certificação energética. Considerando os 289 registos com informação nesta variável, constata-se, pela observação do diagrama de barras produzido, Figura 2, que a classe C (valor 5) é a que ocorre com maior frequência.

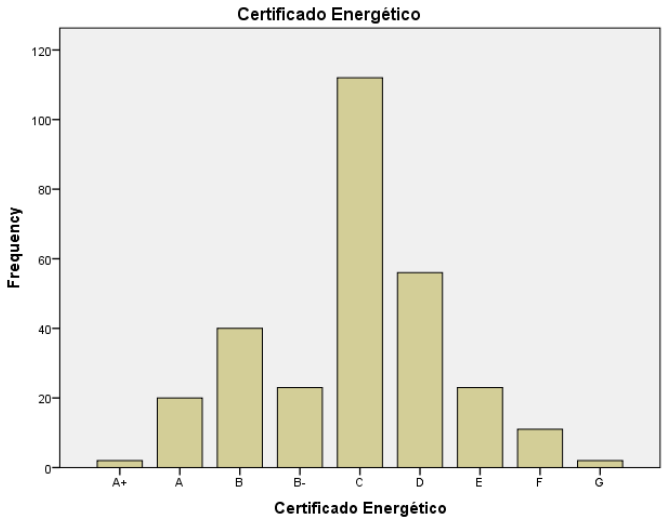

*Figura 2- Diagrama de barras da variável **Energy**.*

Prossegue-se com a análise descritiva das variáveis quantitativas. O Quadro 3 reúne informação respeitante à variável **Price**. Para a amostra recolhida destacam-se o valor médio do preço de arrendamento de um imóvel que é, aproximadamente, 1004 Euros, assim como os preços mínimo e máximo de arrendamento atribuídos a um imóvel, respetivamente, 300 Euros e 111000 Euros. O valor central do preço, ou seja, a mediana do preço, é de 700 Euros. Portanto, 50% dos valores do preço são inferiores ou iguais a 700 Euros e 50% dos valores do preço são superiores ou iguais a 700 Euros. Ainda sobre medidas de localização verifica-se que o 1º quartil ($Q1$) e o 3º quartil ($Q3$) são iguais a 550 e 850 Euros, respetivamente. Por outro lado, o desvio padrão é, aproximadamente, 4781 Euros.

N	Valid	760
	Missing	0
Mean		1003,637
Median		700,000
Std. Deviation		4780,6313
Skewness		20,510
Std. Error of Skewness		,089
Kurtosis		435,611
Std. Error of Kurtosis		,177
Range		110700,0
Minimum		300,0
Maximum		111000,0
Percentiles	25	550,000
	50	700,000
	75	850,000

*Quadro 3- Estatísticas da variável **Price**.*

Para a investigação da presença de valores que se afastam da generalidade das observações recorreu-se à visualização do diagrama de extremos e quartis da variável **Price**, Figura 3. Nesta amostra são detetados dois valores, representados no diagrama pelo símbolo $*$, que se distanciam nitidamente da maioria dos restantes. São identificados como extremos os registos nas posições 251 e 44 da base de dados. Neste caso, extremos são valores superiores a $Q3 + 3 \times AIQ$, sendo AIQ a amplitude interquartis ($Q3 - Q1$).

Foi então efetuada uma avaliação cuidadosa aos dois imóveis, tendo-se verificado que correspondem a dois apartamentos usados, de tipologias $T0$ e $T1$, com preço mensal de arrendamento de 72500 Euros e 111000 Euros, respetivamente. Uma vez que as suas características e descrição apontam claramente no sentido de ter ocorrido um erro de inserção no tipo de negócio por parte do utilizador, ou seja, estes imóveis deveriam ter sido colocados para venda e não arrendamento, optou-se por eliminá-los do estudo.

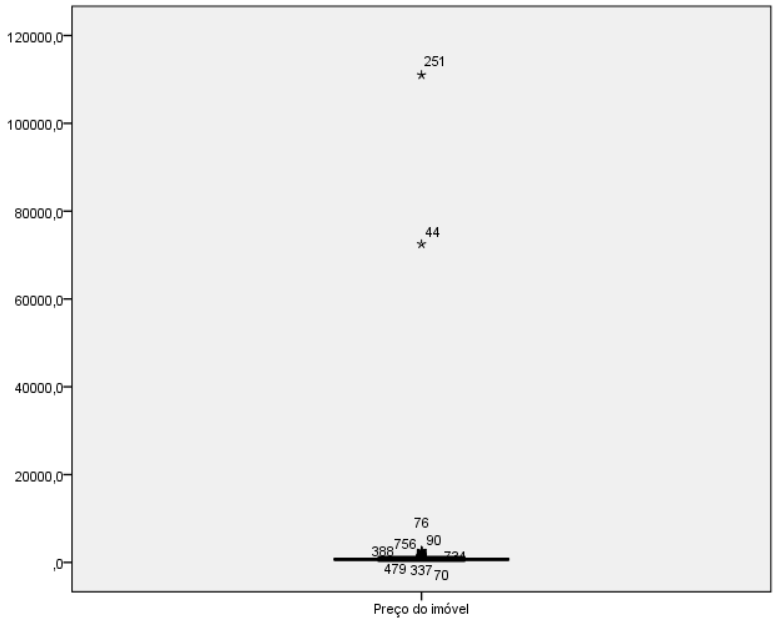

*Figura 3- Diagrama de extremos e quartis da variável **Price**.*

Assim sendo, a base de dados sobre a qual incidirá este trabalho inclui 758 imoveis, dos quais 12 são moradias e 746 são apartamentos. A remoção destes dois imóveis não afeta as descrições feitas atrás para as distribuições das variáveis **Freguesia** e **Energy**. A partir deste momento a análise descritiva assentará na base de dados com 758 imóveis. Volta-se um pouco atrás para novamente aferir sobre a "nova" distribuição da variável **Price**. Segundo o Quadro 4, a média dos preços é agora, aproximadamente, 764 Euros, a mediana 700 Euros e o desvio padrão, aproximadamente 349 Euros. O valor mínimo continua a ser 300 Euros, já o máximo registado foi agora 2700 Euros. Os primeiro, segundo e terceiro quartis mantêm-se inalterados.

N	Valid	758
	Missing	0
Mean		764,201
Median		700,000
Std. Deviation		348,8628
Skewness		2,691
Std. Error of Skewness		,089
Kurtosis		9,863
Std. Error of Kurtosis		,177
Range		2400,0
Minimum		300,0
Maximum		2700,0
Percentiles	25	550,000
	50	700,000
	75	850,000

*Quadro 4- Estatísticas da variável **Price** da amostra de 758 imóveis.*

O diagrama de extremos e quartis, Figura 4, permite visualizar rapidamente não só valores atípicos, mas também dá informação sobre a disposição das medidas de localização e dispersão acabadas de referir. Para além dos extremos (valores superiores a $Q3 + 3 \times AIQ$), são igualmente detetados *outliers* (valores superiores a $Q3 + 1.5 \times AIQ$ mas inferiores ou iguais a $Q3 + 3 \times AIQ$), representados pelo símbolo ∘.

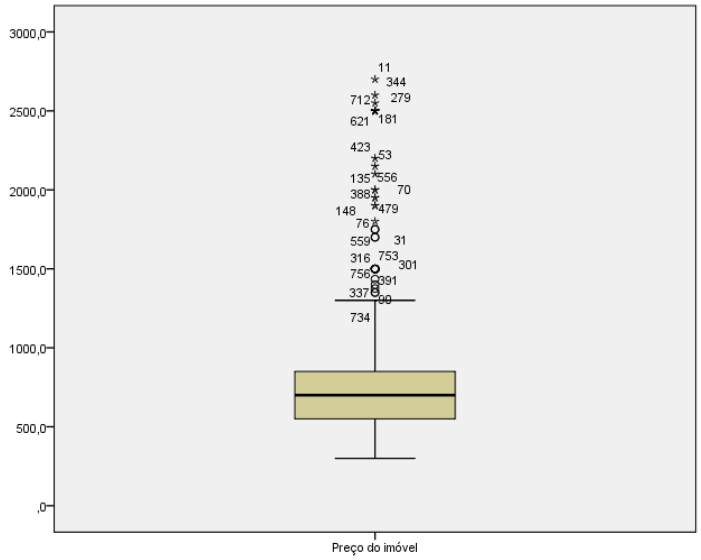

*Figura 4- Diagrama de extremos e quartis da variável **Price** da amostra de 758 imóveis.*

Da Figura 4 e da Figura 5, facilmente se percebe que a distribuição dos preços de arrendamento é assimétrica à direita, facto que pode ser confirmado considerando o coeficiente de assimetria (valor positivo).

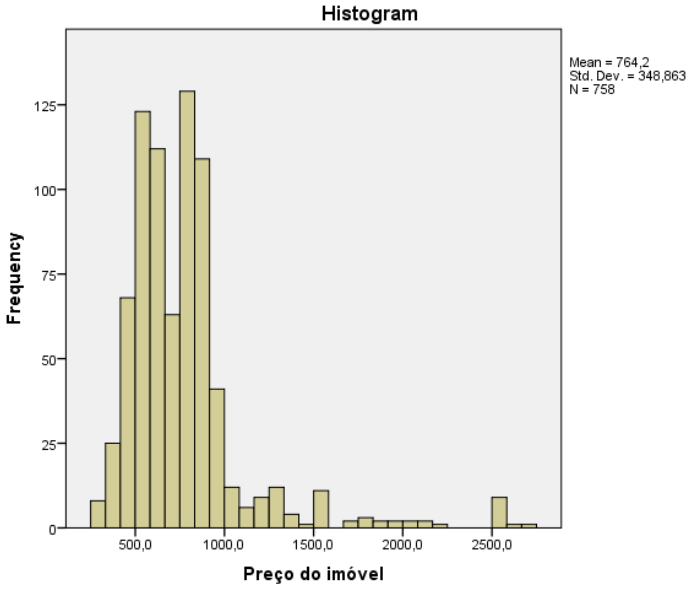

*Figura 5- Histograma da variável **Price** da amostra de 758 imóveis.*

Sendo, tal como a variável **Price**, um dos campos de inserção obrigatória no portal, a variável **UsefulArea** não tem valores omissos. Do Quadro 5, verificamos que o valor esperado é de, aproximadamente, 81.5 m^2 e o desvio padrão de 37.9 m^2.

N	Valid	758
	Missing	0
Mean		81,4974
Median		72,5000
Std. Deviation		37,87415
Skewness		1,590
Std. Error of Skewness		,089
Kurtosis		4,436
Std. Error of Kurtosis		,177
Range		290,00
Minimum		18,00
Maximum		308,00
Percentiles	25	55,0000
	50	72,5000
	75	100,0000

*Quadro 5- Estatísticas da variável **UsefulArea**.*

Relativamente à existência de *outliers* e extremos voltamos a verificar a presença de ambos. À semelhança do que foi feito anteriormente para a variável **Price**, apresenta-se o diagrama de extremos e quartis, Figura 6, onde podemos identificar algumas características da distribuição da variável, tais como o 1º quartil (55 m^2), 2º quartil (72 m^2) e 3º quartil (100 m^2), o máximo (308 m^2)

e mínimo (18 m^2) da amostra. A assimetria é positiva pois, por observação do mesmo diagrama, podemos constatar uma maior concentração de observações nos valores mais reduzidos da amostra.

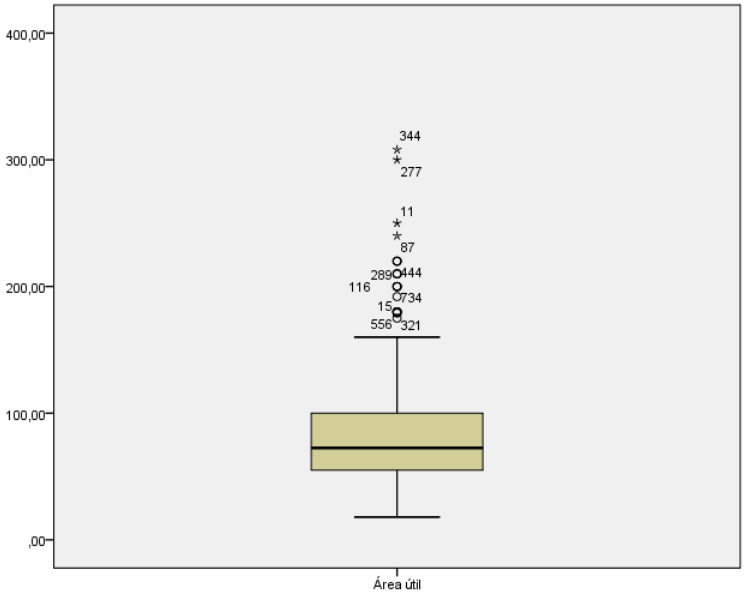

*Figura 6- Diagrama de extremos e quartis da variável **UsefulArea**.*

No que diz respeito à variável **TotalArea**, apenas 232 imóveis apresentam registo neste campo, ou seja, há 526 valores omissos para esta variável. No Quadro 6, à semelhança da análise realizada para a variável **UsefulArea**, apresentam-se algumas estatísticas. Destacam-se: o 1º quartil, 65 m^2, o 2º quartil, 85 m^2, o 3º quartil, 114.75 m^2, o máximo, 262 m^2 e o mínimo, 20 m^2.

N	Valid	232
	Missing	526
Mean		94,0862
Median		85,0000
Std. Deviation		43,92508
Skewness		1,359
Std. Error of Skewness		,160
Kurtosis		2,370
Std. Error of Kurtosis		,318
Range		242,00
Minimum		20,00
Maximum		262,00
Percentiles	25	65,0000
	50	85,0000
	75	114,7500

*Quadro 6- Estatísticas da variável **TotalArea**.*

Pela observação do respetivo diagrama de extremos e quartis, Figura 7, percebemos que a distribuição da variável apresenta igualmente assimetria positiva.

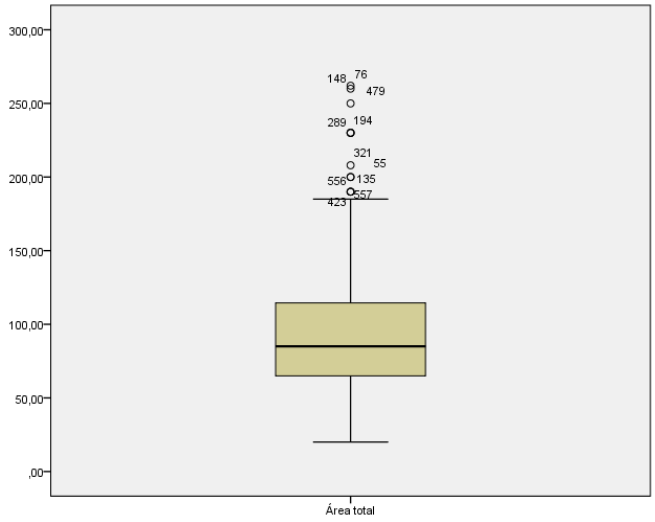

*Figura 7- Diagrama de extremos e quartis da variável **TotalArea**.*

Para tratar a informação relativa à idade de um imóvel recorreu-se à criação da variável **Age** (quantitativa discreta), como já foi referido. Esta variável conta com 582 registos sem informação pelo que apenas 176 são válidos. A idade média de um imóvel é, aproximadamente, 31.40 anos e o desvio padrão é, aproximadamente, 26.35 anos. A distribuição de frequências é assimétrica à direita, predominando imóveis com 1 ano de idade. Assinala-se ainda a presença de um *outlier*. Esta informação pode ser consultada no Quadro 7, na Figura 8 e na Figura 9.

N	Valid	176
	Missing	582
Mean		31,40
Median		30,00
Std. Deviation		26,353
Skewness		,607
Std. Error of Skewness		,183
Kurtosis		-,394
Std. Error of Kurtosis		,364
Range		114
Minimum		1
Maximum		115
Percentiles	25	8,00
	50	30,00
	75	48,50

*Quadro 7- Estatísticas da variável **Age**.*

*Figura 8- Diagrama de extremos e quartis da variável **Age**.*

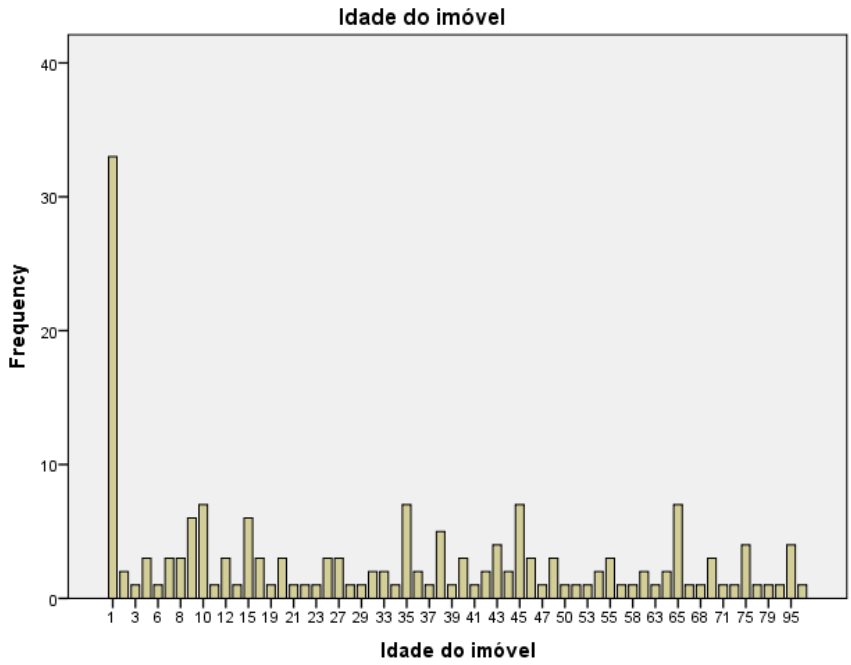

*Figura 9- Diagrama de barras da variável **Age**.*

Prosseguindo com a análise descritiva de **NrRooms**, a variável indica, como o nome sugere, o número de quartos, ou, como já se referiu anteriormente, a tipologia do imóvel. Por se encontrar nas mesmas condições que as variáveis **Price** e **UsefulArea**, ou seja, de preenchimento obrigatório, não apresenta valores omissos. Por observação do Quadro 8, concluímos que a média do número de quartos de um imóvel é, aproximadamente, 2.04 e o valor central da distribuição é 2 quartos. O menor e maior número de quartos observados foram zero e dez, correspondendo a um imóvel do tipo T0 e T10, respetivamente.

N	Valid	758
	Missing	0
Mean		2,04
Median		2,00
Std. Deviation		1,279
Skewness		1,327
Std. Error of Skewness		,089
Kurtosis		3,728
Std. Error of Kurtosis		,177
Range		10
Minimum		0
Maximum		10
Percentiles	25	1,00
	50	2,00
	75	3,00

*Quadro 8- Estatísticas da variável **NrRooms**.*

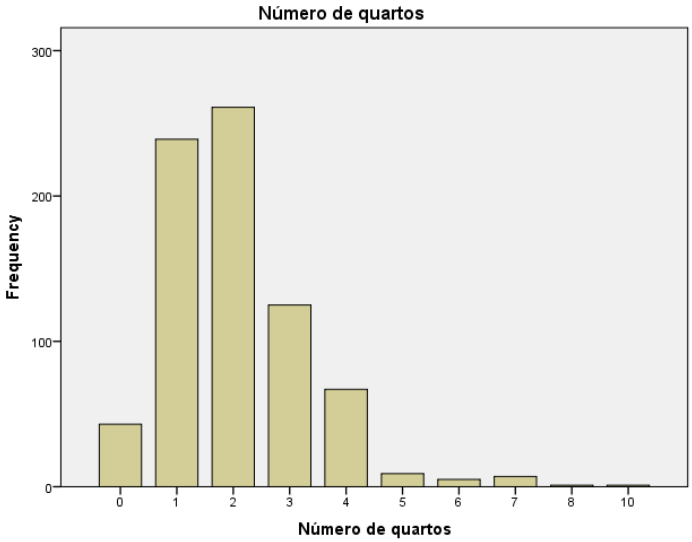

*Figura 10- Diagrama de barras da variável **NrRooms**.*

Pela observação do diagrama de barras, Figura 10, percebemos que a maioria dos imóveis são do tipo T2 seguindo-se a tipologia T1. A distribuição do número de quartos é assimétrica à direita. No que diz respeito à variável **NrWCs** (quantitativa discreta), pela análise do respetivo valor do coeficiente de assimetria, podemos tirar uma conclusão idêntica, ou seja, que a distribuição é também assimétrica positiva. Neste caso, o número médio de casas de banho é, aproximadamente, 1.38. Este valor não suscita desconfiança se tivermos em conta as tipologias mais frequentes registadas atrás. Para 143 registos não foi possível conhecer o número de casas de banho. Estas e outras estatísticas encontram-se no Quadro 9.

N	Valid	615
	Missing	143
Mean		1,38
Median		1,00
Std. Deviation		,602
Skewness		1,574
Std. Error of Skewness		,099
Kurtosis		2,430
Std. Error of Kurtosis		,197
Range		3
Minimum		1
Maximum		4
Percentiles	25	1,00
	50	1,00
	75	2,00

*Quadro 9- Estatísticas da variável **NrWCs**.*

A nova variável, **soma_caracteristicas,** criada com base em informação inserida pelo utilizador do portal, não apresenta valores omissos pelo motivo já exposto, ou seja, pelo facto de se tratar de 23 características que acrescentam valor a um imóvel. Admite-se que a sua não seleção indica que o imóvel não as possui. A média do número de características que valorizam um imóvel é de 2.20 e o número máximo de características registadas num imóvel foi 15. Estas estatísticas, a par de outras, são apresentadas no Quadro 10. Examinando o respetivo diagrama de barras, Figura 11, constata-se que predominam os imóveis com zero características.

N	Valid	758
	Missing	0
Mean		2,20
Median		2,00
Std. Deviation		2,571
Skewness		1,414
Std. Error of Skewness		,089
Kurtosis		2,220
Std. Error of Kurtosis		,177
Range		15
Minimum		0
Maximum		15
Percentiles	25	,00
	50	2,00
	75	4,00

*Quadro 10- Estatísticas da variável **soma_caracteristicas**.*

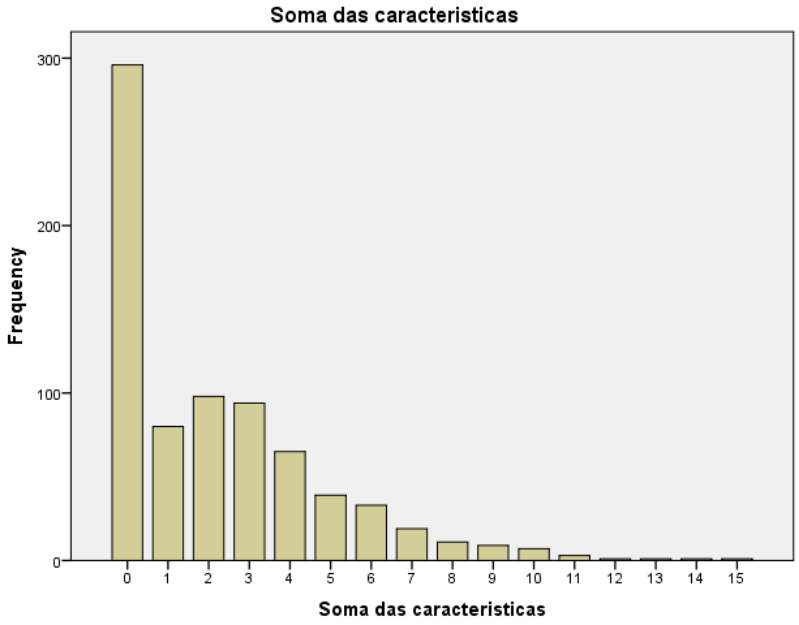

*Figura 11- Diagrama de barras da variável **soma_caracteristicas**.*

Recordando que a variável **NrInteresse** contabiliza os serviços e amenidades existentes na vizinhança de um imóvel, num raio de $1\,Km$, na amostra recolhida verificou-se que o valor esperado é 6.22 pontos de interesse. Segundo a informação constante do Quadro 11, há imóveis que não contabilizam qualquer ponto de interesse neste mesmo raio, sendo, portanto, zero o valor mínimo observado. Pelo contrário, há registo de imóveis com um máximo de 11 pontos de interesse num raio de $1\,Km$. Pelo menos 50% dos imóveis possuem 7 ou menos pontos de interesse no raio referido, ou seja, a mediana da distribuição é 7. Predominam os imóveis com sete pontos de interesse num raio de $1\,Km$, ou seja, a moda é 7. O diagrama de barras da Figura 12 ilustra parte da informação acabada de expor.

N	Valid	758
	Missing	0
Mean		6,22
Median		7,00
Mode		7
Std. Deviation		2,630
Minimum		0
Maximum		11

*Quadro 11- Estatísticas da variável **NrInteresse**.*

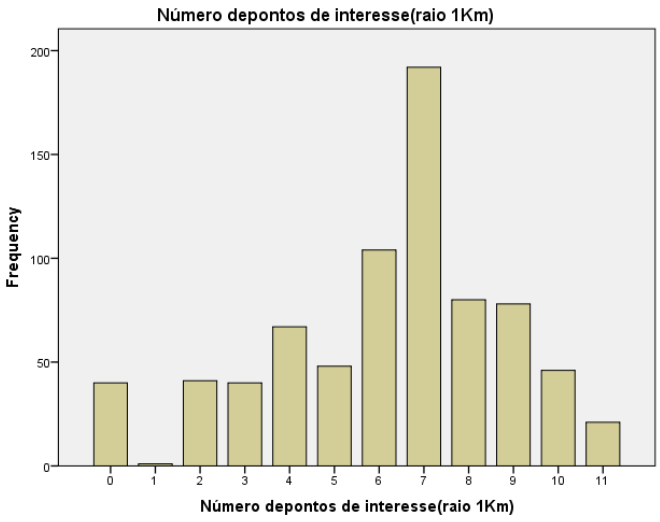

*Figura 12-Diagrama de barras da variável **NrInteresse**.*

Relativamente às variáveis respeitantes aos tempos (em minutos) que demora a percorrer (de carro) as distâncias entre cada imóvel e os pontos de interesse mais próximos já referidos e a

refinaria química, o Quadro 12 e o Quadro 13 compilam as estatísticas amostrais obtidas mais relevantes.

		Tempo ao autocarro (minutos)	Tempo ao café (minutos)	Tempo à mercearia (minutos)	Tempo ao ginásio (minutos)	Tempo ao hospital (minutos)	Tempo ao estabelecimento de diversão noturna (minutos)	Tempo ao parque (minutos)
N	Valid	754	750	748	753	753	754	737
	Missing	4	8	10	5	5	4	21
Mean		1,250973	1,173978	1,519563	2,269079	2,564985	3,592286	2,020556
Median		1,000000	1,000000	1,000000	2,000000	3,000000	3,000000	2,000000
Std. Deviation		1,3320514	1,5231257	1,6024609	1,5361552	1,7790553	2,3395028	1,5486061
Skewness		1,134	1,466	1,811	1,095	,727	,681	1,151
Std. Error of Skewness		,089	,089	,089	,089	,089	,089	,090
Kurtosis		,882	1,289	3,763	2,038	,251	-,144	2,048
Std. Error of Kurtosis		,178	,178	,179	,178	,178	,178	,180
Minimum		,0000	,0000	,0000	,0000	,0000	,0000	,0000
Maximum		5,0000	6,0000	7,0000	7,0000	8,0000	11,5167	9,7500
Percentiles	25	,000000	,000000	,000000	1,000000	1,000000	2,000000	1,000000
	50	1,000000	1,000000	1,000000	2,000000	3,000000	3,000000	2,000000
	75	2,000000	1,954167	2,000000	3,000000	3,000000	5,000000	3,000000

Quadro 12- Estatísticas das variáveis respeitantes a tempos.

		Tempo à farmácia (minutos)	Tempo à refinaria (minutos)	Tempo ao restaurante (minutos)	Tempo à escola (minutos)	Tempo ao metro (minutos)	Tempo ao aterro (minutos)	Tempo à zona industrial (minutos)
N	Valid	753	758	752	752	724	757	756
	Missing	5	0	6	6	34	1	2
Mean		1,503298	12,313500	1,213209	1,350798	3,234438	10,775121	7,871230
Median		1,000000	12,000000	1,000000	1,000000	3,000000	11,000000	8,000000
Std. Deviation		1,4261007	3,1765609	1,4338686	,9983590	2,3265492	2,8013240	2,2986462
Skewness		1,176	3,748	1,364	,608	1,284	-,026	-,596
Std. Error of Skewness		,089	,089	,089	,089	,091	,089	,089
Kurtosis		1,025	55,389	1,186	,036	1,448	-,765	-,038
Std. Error of Kurtosis		,178	,177	,178	,178	,181	,177	,178
Minimum		,0000	6,0000	,0000	,0000	,0000	3,0000	2,0000
Maximum		6,0000	58,0000	5,3500	5,0000	11,0000	16,7667	13,0000
Percentiles	25	,225000	10,000000	,000000	1,000000	1,837500	8,166667	6,000000
	50	1,000000	12,000000	1,000000	1,000000	3,000000	11,000000	8,000000
	75	2,000000	14,391667	2,000000	2,000000	4,000000	13,000000	10,000000

Quadro 13- Estatísticas das variáveis respeitantes a tempos (continuação).

A análise dos histogramas não conduziu ao apontamento de valores que possam ser considerados estranhos. Faz-se notar que esta avaliação pode não ser correta; no entanto, a falta de conhecimento físico da zona envolvida no estudo foi uma dificuldade impossível de ultrapassar. Apenas as variáveis **Time_aterro** e **Time_ZI** apresentam assimetria negativa.

Tentando perceber como se distribuem e associam algumas variáveis, conclui-se esta análise descritiva com uma análise bivariada. Para tal recorremos a diagramas de extremos e quartis, a diagramas de dispersão e coeficientes de correlação. Esta pesquisa baseia-se na intuição e senso comum do investigador, uma vez que não foi possível contar com a ajuda e conhecimento de pessoal especializado em mercado imobiliário, conforme havia sido inicialmente pensado.

Tendo em consideração os diagramas de dispersão e os diagramas de extremos e quartis apresentados a seguir, podemos retirar as seguintes conclusões:

- os maiores valores do preço de arrendamento são, em geral, observados em imóveis do tipo apartamento (ver Figura 13);
- os valores de arrendamento mais baixos são observados, em geral, para imóveis do tipo usado (ver Figura 14);
- como seria de esperar, o preço e a área útil variam no mesmo sentido, ou seja, o preço aumenta quando a área útil aumenta e diminui quando a área útil diminui (ver Figura 15);
- com o aumento da idade do imóvel, o preço de arrendamento tende a diminuir (ver Figura 16);
- observa-se que o preço de arrendamento tende a ser mais elevado à medida que aumenta a tipologia do imóvel; de facto, observando a mediana dos preços dos imóveis de cada tipologia contata-se que tende a ser mais elevada com o aumento do número de quartos (ver Figura 17);
- um aumento no número de casas de banho é acompanhado de um aumento no preço de arrendamento, como seria expectável (ver Figura 18).

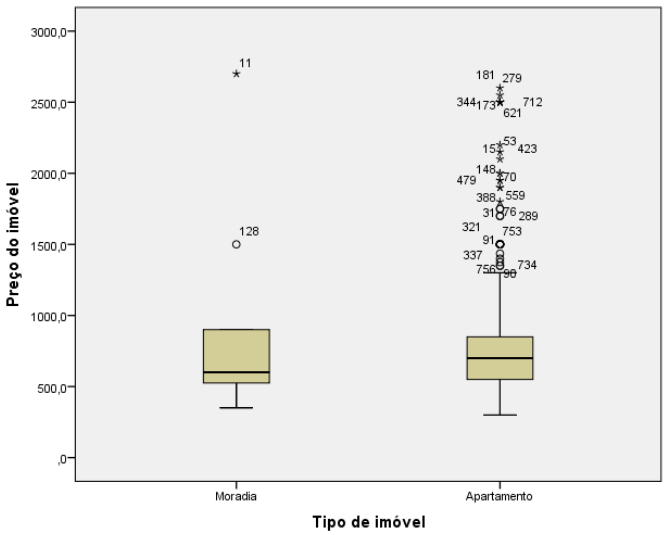

*Figura 13- Diagrama de extremos e quartis para a variável **Price** com os dois tipos de imóveis.*

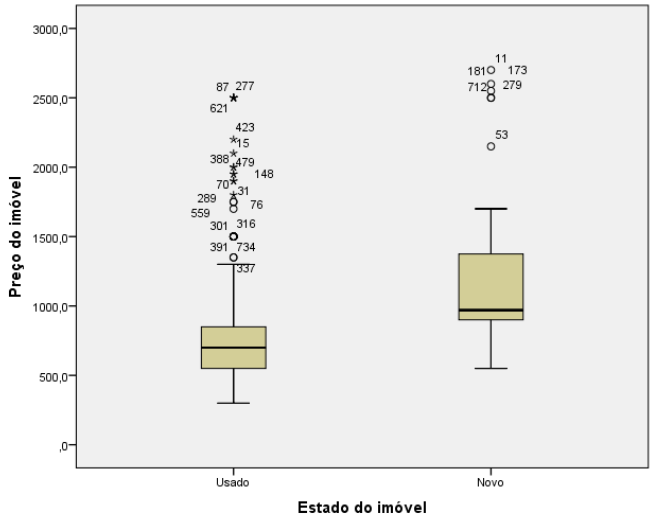

*Figura 14- Diagrama de extremos e quartis para a variável **Price** com os dois tipos de condição.*

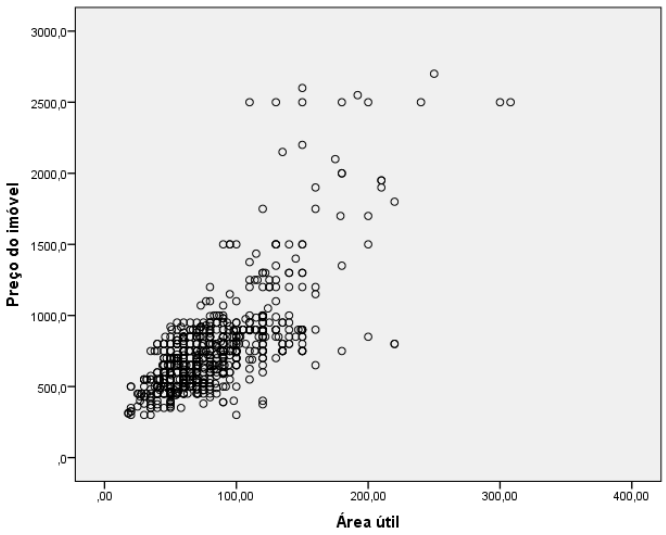

*Figura 15-Diagrama de dispersão entre as variáveis **Price** e **UsefulArea**.*

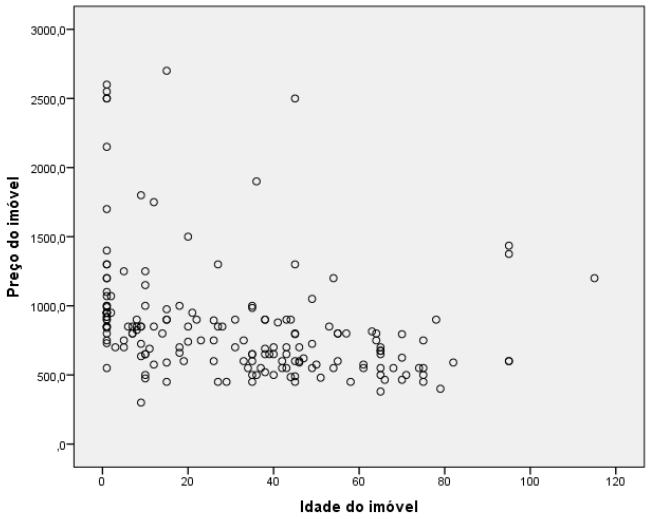

*Figura 16- Diagrama de dispersão entre as variáveis **Price** e **Age**.*

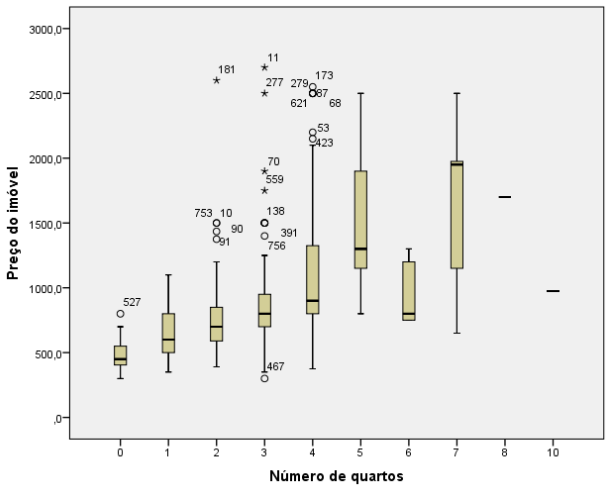

*Figura 17- Diagrama de extremos e quartis para a variável **Price** com as diferentes tipologias.*

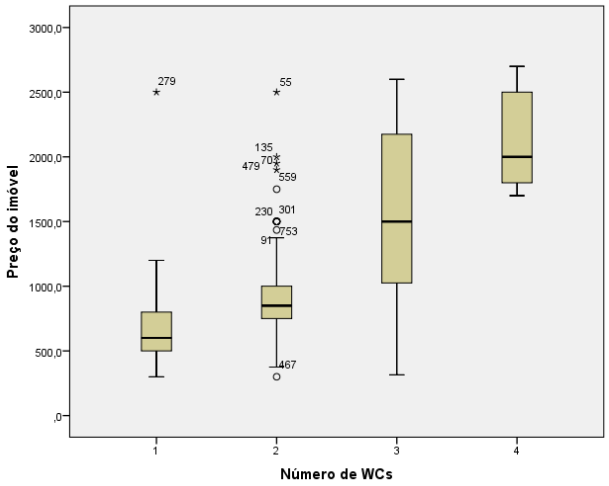

*Figura 18- Diagrama de extremos e quartis para a variável **Price** com o número de casas de banho.*

Finalmente, na Tabela 4, apresentam-se os coeficientes de correlação de Spearman que avaliam o grau de associação entre diferentes variáveis quantitativas (com a exceção das variáveis respeitantes a tempos) e a variável **Energy.** A escolha deste coeficiente deve-se não só ao envolvimento de variáveis independentes do tipo ordinal, **Energy**, mas também ao facto de se tratar de uma medida não paramétrica, pelo que não se torna necessário conhecer a forma da distribuição das variáveis em causa (11). Este coeficiente mede o grau de associação entre duas variáveis, tomando o valor 1 quando a associação é direta perfeita e -1 quando é inversa perfeita. Se as variáveis não estão associadas este coeficiente toma um valor próximo de zero. Entenda-se a expressão "graus de associação" como formas gerais de relacionamento entre as variáveis, que vão além da tão conhecida relação linear (14).

Tabela 4- Coeficientes de correlação de Spearman.

			Preço do imóvel	Número de quartos	Área útil	Certificado Energético	Número de WCs	Soma das caracteristicas	Número de pontos de interesse(raio 1Km)	Área total	Idade do imóvel
Spearman's rho	Preço do imóvel	Correlation Coefficient	1,000	,500**	,672**	-,359**	,539**	,348**	-,050	,670**	-,486**
		Sig. (2-tailed)		,000	,000	,000	,000	,000	,172	,000	,000
		N	758	758	758	288	615	758	758	232	176
	Número de quartos	Correlation Coefficient	,500**	1,000	,822**	,035	,640**	-,049	-,062	,765**	,017
		Sig. (2-tailed)	,000		,000	,558	,000	,174	,088	,000	,824
		N	758	758	758	288	615	758	758	232	176
	Área útil	Correlation Coefficient	,672**	,822**	1,000	-,161**	,684**	,088*	-,097**	,953**	-,246**
		Sig. (2-tailed)	,000	,000		,006	,000	,016	,008	,000	,001
		N	758	758	758	288	615	758	758	232	176
	Certificado Energético	Correlation Coefficient	-,359**	,035	-,161**	1,000	-,173**	-,431**	,056	-,100	,629**
		Sig. (2-tailed)	,000	,558	,006		,005	,000	,342	,368	,000
		N	288	288	288	288	265	288	288	84	55
	Número de WCs	Correlation Coefficient	,539**	,640**	,684**	-,173**	1,000	,144**	-,046	,702**	-,220**
		Sig. (2-tailed)	,000	,000	,000	,005		,000	,260	,000	,006
		N	615	615	615	265	615	615	615	205	152
	Soma das caracteristicas	Correlation Coefficient	,348**	-,049	,088*	-,431**	,144**	1,000	,024	,102	-,419**
		Sig. (2-tailed)	,000	,174	,016	,000	,000		,511	,122	,000
		N	758	758	758	288	615	758	758	232	176
	Número de pontos de interesse(raio 1Km)	Correlation Coefficient	-,050	-,062	-,097**	,056	-,046	,024	1,000	-,052	,292**
		Sig. (2-tailed)	,172	,088	,008	,342	,260	,511		,434	,000
		N	758	758	758	288	615	758	758	232	176
	Área total	Correlation Coefficient	,670**	,765**	,953**	-,100	,702**	,102	-,052	1,000	-,168
		Sig. (2-tailed)	,000	,000	,000	,368	,000	,122	,434		,091
		N	232	232	232	84	205	232	232	232	102
	Idade do imóvel	Correlation Coefficient	-,486**	,017	-,246**	,629**	-,220**	-,419**	,292**	-,168	1,000
		Sig. (2-tailed)	,000	,824	,001	,000	,006	,000	,000	,091	
		N	176	176	176	55	152	176	176	102	176

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Os coeficientes confirmam as conclusões retiradas dos diagramas de dispersão e de extremos e quartis anteriormente elaborados. Restringindo a análise da Tabela 4 ao grau de associação entre o preço do imóvel e as variáveis mencionadas, destaca-se:

- a associação positiva e estatisticamente significativa da variável preço do imóvel com o número de quartos, com a área útil, com o número de casas de banho, com a soma de características e com a área total do imóvel;
- a associação negativa e estatisticamente significativa da variável preço do imóvel com o certificado energético e a idade do imóvel.

As associações verificadas nesta amostra entre as variáveis referidas possuem o sinal esperado. Pela sua pertinência e relevância destacam-se, ainda, as associações estatisticamente significativas entre as variáveis **NrRooms** e **NrWCs, Energy** e **Age**, **soma_caracteristicas** e **Age** e, finalmente, **Energy** e **soma_caracteristicas**.

4. Modelação

Com esta etapa pretende-se desenvolver uma pesquisa exploratória das variáveis que conduzem ao melhor modelo de regressão. Embora a experiência pessoal e intuição nos levem a suspeitar de quais sejam as variáveis mais prováveis de ser incluídas no modelo ajustado final, recorreu-se aos métodos de seleção de variáveis para efetuar esta identificação.

Antes de prosseguir com a investigação da melhor combinação de variáveis independentes que explicam o comportamento médio do preço de arrendamento dos imóveis no concelho de Lisboa, bem como a apresentação e análise geral de alguns modelos de regressão linear, fazem-se duas considerações importantes:

- não serão tidas em conta as variáveis quantitativas **Age**, **TotalArea** e **Energy** por apresentarem valores omissos em elevado número, mais concretamente 582, 526 e 470, respetivamente. A inclusão das variáveis acima referidas levaria à redução da dimensão da base de dados para menos de metade, pelos motivos que se explicam a seguir. Após alguma pesquisa, optou-se por considerar no SPSS Statistics a opção *listwise* em detrimento da *pairwise* pois, com esta preferência, apenas são considerados os imóveis com informação completa, ou seja, com informação em todas as variáveis envolvidas na regressão (15);
- o nível de significância estatístico dos testes de hipóteses efetuados é $\alpha = 0.05$.

As duas primeiras experiências de modelação do problema consideram, para além das quantitativas, apenas uma variável explicativa de natureza qualitativa. Assim, a primeira inclui a variável qualitativa **HouseType** e a segunda a variável qualitativa **State**. Recorda-se que ambos os fatores qualitativos assumem apenas duas modalidades: a **HouseType** assume o valor 0 num imóvel do tipo moradia e o valor 1 se for do tipo apartamento; **State** assume os valores 0 e 1, respetivamente, se o imóvel for usado ou novo. Nestes modelos assume-se que as variáveis qualitativas têm efeito somente no termo independente e que não há interação com outras variáveis independentes.

A terceira experiência envolverá, para além das variáveis quantitativas, as duas variáveis explicativas **State** e **HouseType** simultaneamente, assumindo-se, tal como anteriormente, que nenhum dos fatores qualitativos influencia os coeficientes das variáveis independentes e, ainda, que não há interações entre os dois fatores.

Em cada uma das três experiências seguem-se duas abordagens. Na primeira, todas as variáveis são incluídas na regressão usando os três métodos de seleção: s*tepwise*, *backward* e *forward*. Identificam-se, então, quais as variáveis que os três métodos deixam de fora, as quais serão posteriormente eliminadas do estudo. Finalmente, define-se o modelo ajustado às restantes variáveis, com o método de entrada forçada (12). A segunda abordagem, numa fase inicial, força a entrada de todas as variáveis no modelo. Selecionam-se, então, aquelas cujo efeito na variável dependente, **Price**, é significativo; finalmente, efetua-se uma nova regressão envolvendo somente estas variáveis (15).

Iremos obter assim um total de seis modelos. Os *outputs* obtidos para cada modelo de regressão, com informação adicional para consulta, encontram-se no Anexo A.

1. Primeira experiência: quantitativas com a variável qualitativa HouseType

Considerando a primeira abordagem, depois de introduzir todas as variáveis e efetuar as três análises de regressão com os diferentes métodos de seleção (*stepwise, backward* e *forward*), foram eliminadas da análise as variáveis **HouseType**, **Time_bus_station**, **Time_grocery**, **Time_gym**, **Time_night_club**, **Time_park**, **Time_pharmacy** e **Time_ZI**. A Tabela 5 resume as regressões efetuadas.

*Tabela 5- Variáveis selecionadas por cada método de seleção (com a variável qualitativa **HouseType**).*

	Método de seleção		
Variáveis	*stepwise*	*backward*	*forward*
NrRooms		✓	
UsefulArea	✓	✓	✓
NrWCs	✓	✓	✓
soma_caracteristicas	✓	✓	✓
HouseType			
Time_bus_station			
Time_cafe		✓	
Time_grocery			
Time_gym			
Time_hospital	✓	✓	✓
Time_night_club			
Time_park			
Time_pharmacy			
Time_refinaria	✓		✓
Time_restaurant		✓	
Time_school		✓	
Time_subway		✓	
Time_aterro		✓	
Time_ZI			
NrInteresse		✓	

Desta feita prossegue-se com o ajuste do modelo de regressão às variáveis **NrRooms**, **UsefulArea**, **NrWCs**, **soma_caracteristicas**, **Time_cafe**, **Time_hospital**, **Time_refinaria**, **Time_restaurant**, **Time_school**, **Time_subway**, **Time_aterro** e **NrInteresse**, usando como método a entrada forçada. O primeiro modelo, Modelo 1, é então dado por:

$$\widehat{\boldsymbol{Price}} = 182.632 + 20.715 \times \boldsymbol{NrRooms} + 4.668 \times \boldsymbol{UsefulArea} + 92.711 \times \boldsymbol{NrWCs} + 36.706 \times \boldsymbol{soma\_caracteristicas} - 16.436 \times \boldsymbol{Time\_cafe} - 14.989 \times \boldsymbol{Time\_hospital} + 0.375 \times \boldsymbol{Time\_refinaria} + 13.627 \times \boldsymbol{Time\_restaurant} - 20.023 \times \boldsymbol{Time\_school} - 12.544 \times \boldsymbol{Time\_subway} + 11.430 \times \boldsymbol{Time\_aterro} - 12.696 \times \boldsymbol{NrInteresse}.$$

Considerando a segunda abordagem, numa primeira fase, efetua-se uma regressão forçando a entrada de todas as variáveis e identificam-se as que são relevantes, ou seja, aquelas cujos coeficientes de regressão são estatisticamente significativos por apresentarem um $p - value$ associado ao teste de hipóteses $H_0: \beta_j = 0 \; vs. \beta_j \neq 0, \; j = 2, 3, \dots, k$, inferior ou igual ao nível de significância $\alpha = 0.05$, levando à rejeição da hipótese nula (detalhes desta análise serão apresentados posteriormente). São relevantes as variáveis **UsefulArea**, **NrWCs**,

soma_caracteristicas, **Time_aterro** e **NrInteresse**. Uma nova regressão é executada obrigando a entrada destas variáveis. Obtém-se o Modelo 2:

$$\widehat{\boldsymbol{Price}} = 40.591 + 5.135 \times \boldsymbol{UsefulArea} + 105.551 \times \boldsymbol{NrWCs} + 41.866 \times \boldsymbol{soma\_caracteristicas} + 5.187 \times \boldsymbol{Time\_aterro} + 0.668 \times \boldsymbol{NrInteresse}.$$

2. Segunda experiência: quantitativas com a variável qualitativa State

Seguindo o mesmo raciocínio da experiência anterior, foram selecionadas, através dos três métodos de seleção, as variáveis **NrRooms**, **UsefulArea**, **NrWCs**, **soma_caracteristicas**, **State**, **Time_cafe**, **Time_hospital**, **Time_school**, **Time_subway**, **Time_aterro** e **NrInteresse**. A Tabela 6 resume as regressões efetuadas.

*Tabela 6- Variáveis selecionadas por cada método de seleção (com a variável qualitativa **State**).*

	Método de seleção		
Variáveis	*stepwise*	*backward*	*forward*
NrRooms	✓	✓	✓
UsefulArea	✓	✓	✓
NrWCs	✓	✓	✓
soma_caracteristicas	✓	✓	✓
State	✓	✓	✓
Time_bus_station			
Time_cafe		✓	
Time_grocery			
Time_gym			
Time_hospital	✓	✓	✓
Time_night_club			
Time_park			
Time_pharmacy			
Time_refinaria			
Time_restaurant			
Time_school		✓	
Time_subway	✓	✓	✓
Time_aterro	✓	✓	✓
Time_ZI			
NrInteresse	✓	✓	✓

A regressão final conduziu ao Modelo 3:

$$\widehat{\boldsymbol{Price}} = 231.517 + 21.585 \times \boldsymbol{NrRooms} + 4.500 \times \boldsymbol{UsefulArea} + 98.217 \times \boldsymbol{NrWCs} + 34.613 \times \boldsymbol{Soma\_caracteristicas} + 225.963 \times \boldsymbol{State} - 17.062 \times \boldsymbol{Time\_cafe} - 10.829 \times \boldsymbol{Time\_hospital} - 19.690 \times \boldsymbol{Time\_school} - 14.718 \times \boldsymbol{Time\_subway} + 9.913 \times \boldsymbol{Time\_aterro} - 14.905 \times \boldsymbol{NrInteresse}.$$

Considerando a segunda abordagem, uma vez efetuada uma regressão com entrada forçada de todas as variáveis, verifica-se que são estatisticamente significativos os coeficientes associados às variáveis **NrRooms**, **UsefulArea**, **NrWCs**, **soma_caracteristicas**, **State**, **Time_school**, **Time_subway**, **Time_aterro** e **NrInteresse.** Uma segunda regressão, agora considerando somente estas variáveis, conduz ao Modelo 4:

$$\widehat{\boldsymbol{Price}} = 141.148 + 20.860 \times \boldsymbol{NrRooms} + 4.478 \times \boldsymbol{UsefulArea} + 95.941 \times \boldsymbol{NrWCs} + 36.322 \times \boldsymbol{soma\_caracteristicas} + 227.351 \times \boldsymbol{State} - 17.095 \times \boldsymbol{Time\_school} - 18.408 \times \boldsymbol{Time\_subway} + 9.438 \times \boldsymbol{Time\_aterro} - 5.448 \times \boldsymbol{NrInteresse}.$$

3. Terceira experiência: quantitativas com as variáveis qualitativas State e HouseType

A Tabela 7 resume a informação sobre as variáveis selecionadas pelos três métodos.

*Tabela 7- Variáveis selecionadas por cada método de seleção (com as variáveis qualitativas **HouseType** e **State**).*

	Método de seleção		
Variáveis	*stepwise*	*backward*	*forward*
NrRooms	✓	✓	✓
UsefulArea	✓	✓	✓
NrWCs	✓	✓	✓
soma_caracteristicas	✓	✓	✓
State	✓	✓	✓
HouseType		✓	
Time_bus_station			
Time_cafe		✓	
Time_grocery			
Time_gym			
Time_hospital	✓	✓	✓
Time_night_club			
Time_park			
Time_pharmacy			
Time_refinaria			
Time_restaurant			
Time_school		✓	
Time_subway	✓	✓	✓
Time_aterro	✓	✓	✓
Time_ZI			
NrInteresse	✓	✓	✓

O modelo estimado com as variáveis **NrRooms**, **UsefulArea**, **NrWCs**, **soma_caracteristicas**, **State**, **HouseType**, **Time_cafe**, **Time_hospital**, **Time_school**, **Time_subway**, **Time_aterro** e **NrInteresse** é o Modelo 5:

$$\widehat{\boldsymbol{Price}} = 335.178 + 20.788 \times \boldsymbol{NrRooms} + 4.547 \times \boldsymbol{UsefulArea} + 97.229 \times \boldsymbol{NrWCs} + 34.621 \times \boldsymbol{soma\_caracteristicas} + 226.876 \times \boldsymbol{State} - 95.148 \times \boldsymbol{HouseType} - 17.234 \times \boldsymbol{Time\_cafe} - 11.794 \times \boldsymbol{Time\_hospital} - 19.975 \times \boldsymbol{Time\_school} - 15.977 \times \boldsymbol{Time\_subway} + 9.878 \times \boldsymbol{Time\_aterro} - 15.488 \times \boldsymbol{NrInteresse}.$$

Uma vez mais, considerando a segunda abordagem, ao realizar a análise de regressão com todas as variáveis envolvidas, são estatisticamente significativos os coeficientes associados às variáveis **NrRooms**, **UsefulArea**, **NrWCs**, **soma_caracteristicas**, **State**, **Time_school**, **Time_subway**, **Time_aterro** e **NrInteresse.** O modelo estimado com estas variáveis é o Modelo 6:

$$\widehat{\boldsymbol{Price}} = 141.148 + 20.860 \times \boldsymbol{NrRooms} + 4.478 \times \boldsymbol{UsefulArea} + 95.941 \times \boldsymbol{NrWCs} + 36.322 \times \boldsymbol{Soma\_caracteristicas} + 227.351 \times \boldsymbol{State} - 17.095 \times \boldsymbol{Time\_school} - 18.408 \times \boldsymbol{Time\_subway} + 9.438 \times \boldsymbol{Time\_aterro} - 5.448 \times \boldsymbol{NrInteresse}.$$

A Tabela 8 apresenta, de uma forma resumida, a informação respeitante às variáveis envolvidas em cada modelo, bem como os respetivos coeficientes de regressão estimados. Assinala-se, a sombreado, os coeficientes considerados não significativos. Recorda-se que, segundo os testes de hipóteses $H_0: \beta_j = 0 \; vs. \; \beta_j \neq 0 \;\; , j = 2, 3, \dots, k$, sempre que $p - value > 0.05$, não se rejeita a hipótese nula e conclui-se que, com base nesta amostra e ao nível de significância de 5%, o coeficiente em teste não é estatisticamente significativo. Como consequência, a variável a ele associada não é considerada relevante na explicação do valor esperado do preço de arrendamento do imóvel.

Tabela 8- Resumo dos seis modelos.

	Modelos					
Variáveis	**1**	**2**	**3**	**4**	**5**	**6**
NrRooms	20.715		21.585	20.860	20.788	20.860
UsefulArea	4.668	5.135	4.500	4.478	4.547	4.478
NrWCs	92.711	105.551	98.217	95.941	97.229	95.941
soma_caracteristicas	36.706	41.866	34.613	36.322	34.621	36.322
State	---	---	225.963	227.351	226.876	227.351
HouseType			---	---	-95.148	
Time_bus_station						
Time_cafe	-16.436		-17.062		-17.234	
Time_grocery						
Time_gym						
Time_hospital	-14.989		-10.829		-11.794	
Time_night_club						
Time_park						
Time_pharmacy						
Time_refinaria	0.375					
Time_restaurant	13.627					
Time_school	-20.023		-19.690	-17.095	-19.975	-17.095
Time_subway	-12.544		-14.718	-18.408	-15.977	-18.408
Time_aterro	11.430	5.187	9.913	9.438	9.878	9.438
Time_ZI						
NrInteresse	-12.696	0.668	-14.905	-5.448	-15.488	-5.448
$\overline{R^2}$	**0.628**	**0.630**	**0.649**	**0.644**	**0.650**	**0.644**

Uma análise geral permite concluir que a variável **HouseType** apenas é selecionada uma vez para explicar o comportamento da variável **Price** e, mesmo no Modelo 5 em que está presente, não é considerada relevante, pois o seu coeficiente não é estatisticamente significativo (com base nesta amostra e para $\alpha = 5\%$). Se atendermos aos valores do coeficiente de determinação ajustado, $\overline{R^2}$,

dos Modelos 1 e 2, justamente aqueles que resultaram do envolvimento isolado da variável qualitativa **HouseType**, também constatamos que são os mais baixos. Por conseguinte, os Modelos 1 e 2 não serão alvo de análise mais detalhada. A variável **State** encontra-se envolvida, com efeito significativo, em todas as análises de regressão em que foi incluída. Nos Modelos 5 e 6, foi mesmo a única variável qualitativa selecionada.

O modelo com melhor desempenho, Modelo 5, é o que envolve mais características específicas de localização do imóvel, consideradas relevantes no comportamento do preço médio de arrendamento. Neste modelo ajustado, 65% da variabilidade total em **Price** é explicada pelas variáveis independentes nele incluídas.

Sobressai, igualmente, o facto das características físicas de um imóvel, **UsefulArea**, **NrWCs**, **soma_caracteristicas** e **State** serem sistematicamente variáveis relevantes.

A Tabela 9 apresenta, para os Modelos 3, 4, 5 e 6, os coeficientes estandardizados das variáveis neles envolvidas. Estes coeficientes dão a indicação da variação, em número de desvios padrão, que a variável dependente sofre, quando uma variável independente varia um desvio padrão, *ceteris paribus.*[1] Por este motivo podem ser diretamente comparados, permitindo avaliar a importância de cada uma das variáveis independentes dentro de um mesmo modelo (15).

[1] Expressão com origem no latim que traduz o efeito das outras variáveis ser constante.

Tabela 9- Coeficientes de regressão estandardizados.

	Modelos		
Variáveis	**3**	**4/6**	**5**
NrRooms	0.080	0.078	0.077
UsefulArea	0.485	0.483	0.490
NrWCs	0.169	0.165	0.168
soma_caracteristicas	0.252	0.264	0.252
State	0.140	0.141	0.141
HouseType	---	---	-0.036
Time_bus_station			
Time_cafe	-0.072		-0.073
Time_grocery			
Time_gym			
Time_hospital	-0.055		-0.060
Time_night_club			
Time_park			
Time_pharmacy			
Time_refinaria			
Time_restaurant			
Time_school	-0.058	-0.050	-0.058
Time_subway	-0.096	-0.123	-0.105
Time_aterro	0.080	0.077	0.080
Time_ZI			
NrInteresse	-0.112	-0.041	-0.116
$\overline{R^2}$	**0.649**	**0.644**	**0.650**

Por observação dos valores absolutos dos coeficientes estandardizados inferimos que a variável com maior contribuição relativa para explicar o comportamento da variável **Price** é **UsefulArea**. Seguem-se **soma_caracteristicas**, **NrWCs** e **State**. A variável **NrRooms**, embora apareça em todos os modelos, não é relevante em nenhum.

Time_school, **Time_subway**, **Time_aterro** e **NrInteresse** figuram também em todos os modelos. No entanto, analisando a significância dos seus coeficientes, **Time _subway** e **Time_aterro** serão as mais fortes (e relevantes em todos os modelos). Surgem ainda duas variáveis caraterizadoras da vizinhança de um imóvel, **Time_cafe** e **Time_hospital**, ainda que nunca sejam relevantes.

No conjunto das 21 variáveis iniciais podemos então concluir, com base nesta amostra, que há um grupo que não tem influência no comportamento do preço de arrendamento, pois não foram escolhidas pelos métodos de seleção ou, tendo sido escolhidas, os seus coeficientes não foram considerados estatisticamente significativos ao nível de significância de 5%. São elas **NrRooms**,

HouseType, **Time_bus_station**, **Time_cafe**, **Time_grocery**, **Time_gym**, **Time_hospital**, **Time_night_club**, **Time_park**, **Time_pharmacy**, **Time_refinaria**, **Time_restaurant** e **Time_ZI**. Nas restantes variáveis podemos considerar a existência de dois níveis de importância, distinguindo **UsefulArea**, **NrWCs**, **soma_caracteristicas**, **State, Time_subway** e **Time_aterro** por serem aquelas que são recorrentemente selecionadas; num segundo patamar incluímos **NrInteresse** e **Time_school**, pois a sua relevância não é unânime entre modelos.

Dirigindo agora a atenção para os sinais dos coeficientes das variáveis relevantes, apenas **Time_school**, **Time_subway** e **NrInteresse** apresentam sinais negativos. O sinal do coeficiente de regressão associado a **Nrinteresse** não era expectável. Relativamente às restantes variáveis, este sinal fará todo o sentido se pensarmos que um aumento do tempo de deslocação (e consequente aumento da distância) desde o imóvel a uma escola ou estação de metro leva (intuitivamente) à diminuição do preço de arrendamento. No que diz respeito aos sinais positivos dos coeficientes, eles apontam no sentido de serem variáveis cujo aumento induz também um aumento na variável **Price**.

Em suma, nesta análise exploratória tentou-se perceber em que medida a localização de um imóvel, a par das suas características físicas, influencia o comportamento do seu preço de arrendamento ou, dito de outra forma, tentou-se identificar quais os atributos determinantes na formação do preço de arrendamento. Para esse efeito, para além da típica informação estrutural de um imóvel, foram efetuadas algumas medidas de localização do mesmo, que se concretizaram em variáveis de medição de tempos que se demora a percorrer determinadas distâncias desde o imóvel a um ponto de interesse ou local de considerado valor. Após a investigação de alguns modelos de regressão, conclui-se que destas variáveis, apenas duas, **Time_aterro** e **Time_subway**, e, eventualmente, também **Time_school**, contribuem para a variação observada no preço. Da análise da Tabela 9 constata-se que são as variáveis correspondentes a atributos físicos que mais afetam e, portanto, mais importância possuem na explicação do fenómeno em estudo.

A fim de aferir sobre quais as características específicas de um imóvel que podem determinar o seu preço de arrendamento, efetua-se agora uma análise onde se discrimina a presença das características que deram origem à formação da variável **soma_caracteristicas**. Uma análise exploratória semelhante às realizadas anteriormente com a primeira abordagem, envolvendo as variáveis **UsefulArea**, **NrWCs**, **State**, **Time_subway**, **Time_aterro**, **NrInteresse** e as 23 variáveis binárias correspondentes às características específicas, permite obter o Modelo 7 estimado:

$$\widehat{\boldsymbol{Price}} = 83.489 + 5.083 \times \boldsymbol{UsefulArea} + 98.778 \times \boldsymbol{NrWCs} + 217.817 \times \boldsymbol{State} - 14.923 \times \boldsymbol{Time\_subway} + 8.098 \times \boldsymbol{Time\_aterro} + 57.498 \times \boldsymbol{Cozinha\_equipada} + 102.902 \times \boldsymbol{Estacionamento} + 48.174 \times \boldsymbol{Varanda} + 141.411 \times \boldsymbol{Ar\_condicionado} + 143.343 \times \boldsymbol{Condominio\_fechado}.$$

Este modelo apresenta um coeficiente de determinação ajustado, $\overline{R^2}$, de 0.653 e todas as variáveis são consideradas relevantes, com base nesta amostra e ao nível de significância de 5%.

Numa tentativa de conseguir um modelo com melhor desempenho, e atendendo um pouco à sensibilidade pessoal, considerou-se um modelo no qual constam, para além dos atributos físicos e de localização selecionados na análise anterior, mais um atributo de localização específico, **Time_hospital**, e um de caracterização de vizinhança mais geral, **NrInteresse**. Apresenta-se o Modelo 8, estimado:

$$\widehat{\boldsymbol{Price}} = 143.019 + 5.140 \times \boldsymbol{UsefulArea} + 98.994 \times \boldsymbol{NrWCs} + 215.688 \times \boldsymbol{State} - 11.277 \times \boldsymbol{Time\_subway} + 10.247 \times \boldsymbol{Time\_aterro} - 19.523 \times \boldsymbol{Time\_hospital} - 7.667 \times \boldsymbol{NrInteresse} + 56.079 \times \boldsymbol{Cozinha\_equipada} + 99.340 \times \boldsymbol{Estacionamento} + 49.616 \times \boldsymbol{Varanda} + 139.322 \times \boldsymbol{Ar\_condicionado} + 140.374 \times \boldsymbol{Condominio\_fechado}.$$

Com este modelo, 65.8% da variabilidade total da variável dependente é explicada pelas variáveis independentes presentes. Houve, portanto, uma melhoria relativamente ao modelo anterior.

Todas as variáveis, à exceção de **NrInteresse**, contribuem de forma significativa para explicar possíveis variações no preço de arrendamento. Todavia, se elevarmos o nível de significância do teste para 0.1, esta variável torna-se relevante. Quanto às contribuições relativas, tendo como base os coeficientes estandardizados apresentados em anexo, conclui-se que **UsefulArea** mantém o seu posto, seguido de **NrWCs** e **State.**

Excluindo, então, a variável **NrInteresse** obtém-se o seguinte Modelo 9:

$$\widehat{\boldsymbol{Price}} = 93.733 + 5.132 \times \boldsymbol{UsefulArea} + 100.224 \times \boldsymbol{NrWCs} + 213.707 \times \boldsymbol{State} - 10.129 \times \boldsymbol{Time\_subway} + 8.401 \times \boldsymbol{Time\_aterro} - 13.689 \times \boldsymbol{Time\_hospital} + 56.197 \times \boldsymbol{Cozinha\_equipada} + 98.401 \times \boldsymbol{Estacionamento} + 50.454 \times \boldsymbol{Varanda} + 137.894 \times \boldsymbol{Ar\_condicionado} + 146.790 \times \boldsymbol{Condominio\_fechado}.$$

Este modelo apresenta um $\overline{R^2} = 0.656$ e todas as variáveis são relevantes. Prossegue-se com uma análise mais aprofundada do mesmo no capítulo 5.

5. Análise do modelo de regressão linear múltipla selecionado

Tal como foi referido, neste capítulo vai proceder-se à análise detalhada do Modelo 9. Far-se-á um estudo deste modelo no que diz respeito à validação dos seus pressupostos, à análise e interpretação dos coeficientes de regressão, à identificação de *ouliers* e casos influentes, entre outros aspetos.

Recorde-se o Modelo 9:

$$\widehat{\boldsymbol{Price}} = 93.733 + 5.132 \times \boldsymbol{UsefulArea} + 100.224 \times \boldsymbol{NrWCs} + 213.707 \times \boldsymbol{State} - 10.129 \times \boldsymbol{Time\_subway} + 8.401 \times \boldsymbol{Time\_aterro} - 13.689 \times \boldsymbol{Time\_hospital} + 56.197 \times \boldsymbol{Cozinha\_equipada} + 98.401 \times \boldsymbol{Estacionamento} + 50.454 \times \boldsymbol{Varanda} + 137.894 \times \boldsymbol{Ar\_condicionado} + 146.790 \times \boldsymbol{Condominio\_fechado}.$$

Começamos por fazer notar que, da base inicial com 758 imóveis, com a opção *listwise* foram considerados 565 registos para o modelo de regressão ajustado. Nele estão envolvidas onze variáveis independentes, e, portanto, doze coeficientes de regressão ($k = 12$).

Na tabela "Model Summary", Tabela 10, do *output* do SPSS, podemos verificar que $\overline{R^2} = 0.656$, isto é, 65.6% da variabilidade total da variável dependente, **Price**, é explicada pelas variáveis independentes incluídas no modelo estimado.

Tabela 10- Estatísticas do Modelo 9.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,814[a]	,663	,656	202,4018	,663	98,787	11	553	,000	1,862

5.1. Deteção de *outliers* e casos influentes

Uma vez que os valores estimados dos coeficientes de regressão podem ser influenciados pela presença de casos que não seguem a tendência da maioria, vamos investigar a sua existência recorrendo à análise dos resíduos. Se um registo é um *outlier* então o modelo pode não ser capaz de o modelar com muita precisão. Por este motivo, espera-se também que o resíduo correspondente seja elevado, em valor absoluto, e se afaste da maioria das outras observações (15). Através dos resíduos estandardizados podemos efetuar esta avaliação, esperando que em 95% dos casos os resíduos estejam compreendidos entre -1.96 e 1.96, 99% estejam compreendidos entre -2.58 e 2.58 e 99.9% entre -3.29 e 3.29 (regra empírica; assumindo Normalidade). Observando a tabela "Casewise Diagnostics", Tabela 11, constata-se que:

- 28 casos (aproximadamente 5%) têm um resíduo estandardizado com valor absoluto maior que 2;
- aproximadamente 97% do casos (546 casos) estão compreendidos entre -2.58 e 2.58;
- os casos 55, 87, 181, 194, 279, 423 e 619 (aproximadamente 0.1% dos casos) apresentam um resíduo estandardizado com valor absoluto superior a 3.29.

Tabela 11- Imóveis que se afastam mais de 2 desvios padrão.

Casewise Diagnostics

Case Number	Std. Residual	Preço do imóvel	Predicted Value	Residual
5	-2,752	900,0	1457,013	-557,0126
53	3,151	2150,0	1512,184	637,8156
54	-2,048	720,0	1134,420	-414,4204
55	4,837	2500,0	1520,981	979,0186
70	2,918	1900,0	1309,478	590,5217
87	3,420	2500,0	1807,795	692,2048
135	2,817	2000,0	1429,826	570,1739
177	2,051	1300,0	884,793	415,2066
181	4,264	2600,0	1736,896	863,1045
194	-3,333	800,0	1474,603	-674,6028
195	-3,283	800,0	1464,474	-664,4736
212	2,858	1500,0	921,600	578,4000
229	2,877	1500,0	917,745	582,2549
230	2,750	1500,0	943,406	556,5940
264	-2,508	315,0	822,664	-507,6643
279	4,964	2500,0	1495,200	1004,8000
416	-3,211	400,0	1049,905	-649,9046
423	5,003	2200,0	1187,362	1012,6381
426	-2,012	900,0	1307,259	-407,2586
467	-2,172	300,0	739,649	-439,6488
479	2,578	1950,0	1428,225	521,7747
533	-3,008	375,0	983,922	-608,9221
559	3,082	1750,0	1126,215	623,7849
619	4,636	2500,0	1561,633	938,3671
695	-2,128	850,0	1280,651	-430,6509
710	2,292	2500,0	2036,057	463,9429
722	-2,197	750,0	1194,605	-444,6054
730	-2,775	750,0	1311,567	-561,5668

Dada a dimensão da amostra, $n = 565$, as duas últimas constatações não se consideram alarmantes. Esta análise simples não levanta suspeitas relativamente à adequação do modelo ajustado.

Para averiguar a existência de casos influentes no modelo, recorremos à estatística da distância de Cook, que pode ser consultada na tabela "Residual Statistics", Tabela 12. Considerando como valor de corte o valor 1, ou seja, registos com uma distância de Cook superior a 1 são

preocupantes, neste modelo, nunca se verifica esta situação. Portanto, segundo este critério, não há casos que se possam considerar influentes (12).

Tabela 12- Estatísticas dos resíduos.

Residuals Statistics

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	300,618	2167,008	773,080	280,9442	565
Std. Predicted Value	-1,682	4,962	,000	1,000	565
Standard Error of Predicted Value	14,407	67,141	27,967	9,387	565
Adjusted Predicted Value	300,471	2138,403	773,082	279,8150	565
Residual	-674,6028	1012,6381	,0000	200,4183	565
Std. Residual	-3,333	5,003	,000	,990	565
Stud. Residual	-3,415	5,194	,000	1,008	565
Deleted Residual	-708,0074	1100,0493	-,0026	207,8348	565
Stud. Deleted Residual	-3,448	5,321	,001	1,015	565
Mahal. Distance	1,859	61,064	10,981	9,018	565
Cook's Distance	,000	,213	,003	,014	565
Centered Leverage Value	,003	,108	,019	,016	565

Em alternativa, podemos ainda considerar os *DFBetas* que traduzem a diferença entre os parâmetros de regressão estimados usando todos os casos e excluindo uma observação sequencialmente. Olhando para estes valores, é também possível reconhecer os casos que exercem uma grande influência na determinação dos parâmetros do modelo. Como referência investigamos os casos com valor absoluto maior que 2 (16). Na amostra utilizada não se observou nenhuma observação influente segundo este critério.

Continua-se a análise do Modelo 9 com o teste de significância global da regressão que compreende as seguintes hipóteses:

$$H_0: \beta_2 = \beta_3 = \cdots = \beta_{12} = 0 \qquad vs. \qquad H_1: \exists\, \beta_j \neq 0 \ (j = 2,3, \dots, 12).$$

O $p - value$ associado ao teste pode ser lido na tabela da ANOVA da regressão, Tabela 13. Como o $p - value$ é aproximadamente zero, rejeita-se a hipótese nula a favor de H_1. Ou seja, pelo menos uma das variáveis independentes contribui significativamente para explicar o comportamento esperado do preço de arrendamento, uma conclusão que, naturalmente, vem corroborar a análise do Modelo 9 realizada no término do capítulo anterior.

Tabela 13- ANOVA do Modelo 9.

ANOVA

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	44516316,910	11	4046937,901	98,787	,000[b]
	Residual	22654474,506	553	40966,500		
	Total	67170791,416	564			

5.2. Análise dos coeficientes de regressão linear

Apresentam-se agora, em detalhe, os testes realizados anteriormente e que justificam a definição do Modelo 9. Recorda-se que as hipóteses em causa são:

$$H_0: \beta_j = 0 \qquad vs. \qquad H_1: \beta_j \neq 0 \ (j = 1, 2, 3, \dots, 12).$$

Os resultados podem ser consultados na tabela "Coefficients", Tabela 14. Constata-se que todos os coeficientes correspondentes às variáveis independentes são significativos, uma vez que $p - value \leq 0.05$. Portanto, todos os regressores são considerados relevantes para explicar o comportamento médio de **Price**, com base nesta amostra e ao nível de significância especificado.

Os sinais dos coeficientes são negativos para as variáveis **Time_hospital** e **Time_subway**, indicando que um aumento no tempo de deslocação (e, portanto, aumento da distância) desde o imóvel a um destes pontos de interesse leva à diminuição do preço de arrendamento. Os restantes coeficientes apresentam sinal positivo, denunciando uma relação positiva. Como seria de esperar, a área útil e o número de casas de banho são atributos que induzem um aumento na variável dependente. O conjunto de características específicas consideradas no modelo de regressão produzem o mesmo efeito; este cenário era esperado se se tiver em conta que são particularidades que acrescentam valor a um imóvel. O coeficiente associado à distância do imóvel ao aterro aponta no sentido de que quanto mais afastados estão, mais elevado é o preço, situação também expectável. Finalmente, como era esperado, o estado do imóvel também influi positivamente a variável dependente.

A análise comparativa dos valores absolutos dos coeficientes de regressão estandardizados permite concluir que os atributos físicos área útil, número de casas de banho e estado do imóvel são os que apresentam maior contribuição relativa para explicar o comportamento esperado do preço. Seguem-se as variáveis, à exceção de uma, **Varanda**, que assinalam a presença ou ausência de características específicas e, finalmente, as variáveis respeitantes à localização do imóvel.

Com base nesta amostra e no modelo especificado, apresenta-se, com mais detalhe, a interpretação de alguns coeficientes de regressão:

- estima-se que, em média, o aumento de um metro quadrado na variável independente **UsefulArea** induz um aumento de 5.132 Euros na variável **Price**, *ceteris paribus*;
- estima-se que, em média, o aumento de uma casa de banho induz um aumento de 100.24 Euros na variável **Price**, *ceteris paribus*;
- estima-se que, em média, um minuto adicional na variável **Time_subway** induz em **Price** um decréscimo de 10.13 Euros, *ceteris paribus*;
- estima-se que, em média, o aumento de um minuto na variável **Time_aterro** induz em **Price** um aumento de 8.40 Euros, *ceteris paribus*;
- ter cozinha equipada faz aumentar o preço de arrendamento em 56.20 Euros, *ceteris paribus*.

Neste momento revela-se oportuno referir uma interpretação baseada nos intervalos de confiança (IC) dos coeficientes de regressão e que é comummente descurada na literatura. A título de exemplo, considere-se a variável independente **UsefulArea** cujo IC a 95% é dado por $[4.44;\ 5.82]$, aproximadamente. Concluímos que, com base nesta amostra e modelo especificado, em média, o aumento de um metro quadrado na variável independente **UsefulArea** induz um aumento no preço de arrendamento do imóvel que pode oscilar entre 4.44 Euros e 5.82 Euros, *ceteris paribus*.

Tabela 14- Coeficientes de regressão estimados do Modelo 9.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	93,733	48,056		1,951	,052	-,661	188,127					
	Área útil	5,132	,350	,553	14,655	,000	4,444	5,820	,717	,529	,362	,428	2,339
	Estado do imóvel	213,707	42,394	,132	5,041	,000	130,434	296,980	,316	,210	,124	,887	1,128
	Número de WCs	100,224	21,743	,173	4,610	,000	57,515	142,932	,637	,192	,114	,434	2,307
	Tempo ao hospital (minutos)	-13,689	5,588	-,070	-2,450	,015	-24,665	-2,713	-,076	-,104	-,060	,751	1,332
	Tempo ao metro (minutos)	-10,129	4,386	-,068	-2,309	,021	-18,745	-1,514	-,131	-,098	-,057	,705	1,418
	Tempo ao aterro (minutos)	8,401	3,411	,068	2,463	,014	1,700	15,102	-,195	,104	,061	,789	1,267
	Cozinha _equipada	56,197	19,422	,081	2,894	,004	18,048	94,347	,135	,122	,071	,771	1,297
	Estacionamento	98,401	26,234	,110	3,751	,000	46,871	149,931	,294	,158	,093	,707	1,415
	Varanda	50,454	19,446	,069	2,595	,010	12,258	88,651	,266	,110	,064	,855	1,170
	Ar_condicionado	137,894	30,158	,119	4,572	,000	78,656	197,133	,312	,191	,113	,907	1,102
	Condominio_ fechado	146,790	46,022	,084	3,190	,002	56,390	237,190	,172	,134	,079	,877	1,141

5.3. Análise dos pressupostos do modelo

Para averiguar o pressuposto da Normalidade dos erros, o SPSS disponibiliza, aquando da execução da regressão linear, um gráfico de probabilidade Normal cujo eixo das abcissas representa a probabilidade observada acumulada dos resíduos e o eixo das ordenadas a probabilidade acumulada que se observaria se os resíduos seguissem uma distribuição Normal. Este gráfico de probabilidade Normal encontra-se representado na Figura 19.

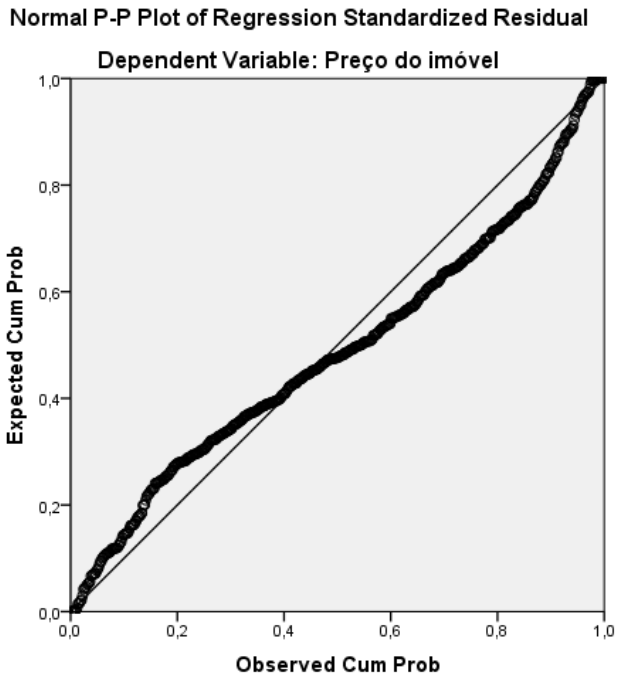

Figura 19- Gráfico de probabilidade Normal.

Da observação da Figura 19, não obstante os desvios observados, o pressuposto de Normalidade dos erros não é invalidado com esta informação dos resíduos.

Ainda neste contexto, efetuou-se um teste de ajustamento dos resíduos a uma distribuição Normal, o teste de Kolmogorov-Smirnov, com correção de Lilliefors, tendo-se obtido $p - value < 0.001$, o que conduz à rejeição da hipótese nula, ou seja, da Normalidade dos erros. Todavia, importa aqui referir a fragilidade deste teste, uma vez que, quando as amostras têm uma dimensão elevada, como é o caso, ele tende a rejeitar indevidamente a hipótese nula com elevada frequência (13). Note-se que, mesmo considerando a ausência de Normalidade, a inferência permanece válida, em termos aproximados, pois este trabalho é baseado numa amostra de grande dimensão.

Segue-se com a análise do pressuposto da homocedasticidade dos resíduos. O gráfico dos resíduos estandardizados *vs.* valores preditos estandardizados, Figura 20, denuncia a possível existência de heteroscedasticidade. Esta situação percebe-se pela distribuição dos pontos do gráfico que, como se observa, não apresenta uma mancha de pontos aleatórios com o mesmo tipo de dispersão em torno do eixo das abcissas.

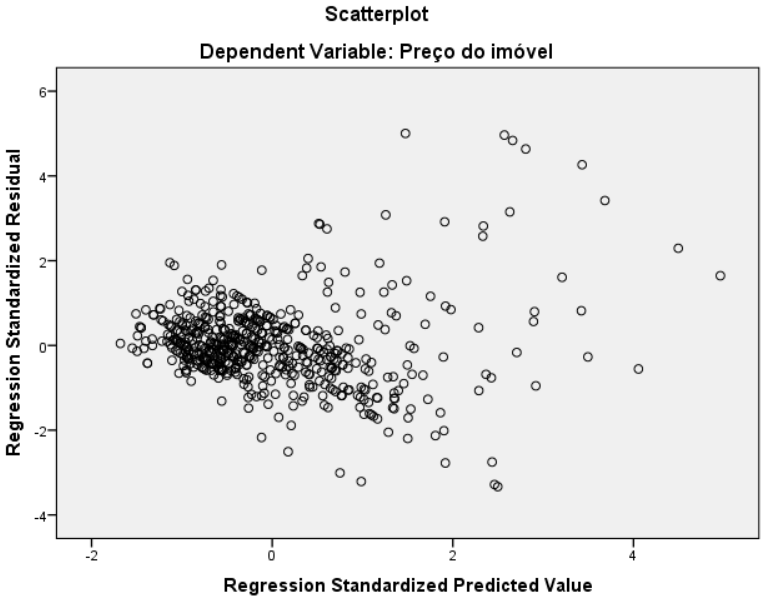

Figura 20- Gráfico dos resíduos estandardizados vs. valores preditos estandardizados.

Para confirmação da suspeita, realizou-se o teste de White simplificado (11):

$$H_0: \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_{565}^2$$

$$vs.$$

$$H_1: \sigma_t^2 \neq \sigma_s^2 \text{ , para algum } t \neq s.$$

O teste engloba os seguintes passos:

1º) calcular os resíduos por mínimos quadrados, $\hat{u}_t = y_t - \hat{y}_t$, $t = 1, \ldots, n$ do modelo (variável automaticamente guardada pelo SPSS);

2º) fazer a regressão auxiliar de $\hat{u}_t^2$ sobre **1**, $\boldsymbol{Price\_predicted}$ (variável automaticamente guardada pelo SPSS) e $\boldsymbol{Price\_predicted^2}$, e retirar o valor do coeficiente de determinação da regressão correspondente, R^2;

3º) calcular o valor da estatística de teste $W_s = nR^2 \sim \chi^2(2)$, onde $n = 565$.

Efetuando este procedimento obteve-se $W_s \approx 127.69$. Notando que a região de rejeição do teste se encontra na aba direita da distribuição, o valor crítico da distribuição Qui-Quadrado com 2

graus de liberdade e $\alpha = 0.05$ é, aproximadamente, 5.99, o que conduz à rejeição da hipótese nula. É importante salientar que, apesar deste pressuposto não se observar, a análise de regressão efetuada não é invalidada, pois a relação linear entre as variáveis não deixa de ser detetada. Simplesmente a análise é enfraquecida (17). Recorde-se que, nestas condições, o estimador dos mínimos quadrados permanece consistente e não enviesado, deixando simplesmente de ser o mais eficiente entre os estimadores lineares não enviesados. Em trabalho futuro sugere-se a aplicação do estimador de White (18).

Prossegue-se com a averiguação do pressuposto de ausência de autocorrelação, recorrendo ao conhecido teste de Durbin-Watson, d. O valor da estatística de teste é 1.862, Tabela 10, ou seja, $d \approx 2.0$ (± 0.4; regra empírica), pelo que não se rejeita a hipótese nula da ausência de autocorrelação entre os erros do modelo (12).

Finalmente avalia-se a associação entre as variáveis independentes presentes no modelo. O primeiro diagnóstico de colinearidade foi feito através da análise do "Fator de inflação da variância", VIF, e pela estatística *Tolerance* $\left(\frac{1}{VIF}\right)$, para cada variável independente (ver Tabela 14). Este fator esclarece sobre a relação linear de uma variável independente com as restantes variáveis independentes. Os valores observados de VIF variam entre, aproximadamente, 1.1 e 2.4. Sendo menores que 10, não são considerados alarmantes (Bowerman e O´Connell, 1990 e Myers, 1990 citados por 15, p. 242). Relativamente aos valores de *Tolerance*, sendo maiores que 0.2 (Menard, 1995 citado por 15, p. 242), não indicam a existência de problemas. Podemos concluir que não há motivos para preocupação com a colinearidade entre as variáveis independentes. O segundo diagnóstico baseou-se nos valores de *Condition Index* constantes da tabela "Colinearity Diagnostics",Tabela 15. Os valores observados, inferiores a 30, não apontam no sentido de haver problemas na estimação dos coeficientes de regressão devido à presença de colinearidade entre as variáveis independentes (Belskey, Kuh e Welsch, 1980 citados por 12, p. 715).

Tabela 15- Estatísticas de diagnóstico de colinearidade.

Collinearity Diagnostics

Model		Eigenvalue	Condition Index	Variance Proportions											
				(Constant)	Área útil	Estado do imóvel	Número de WCs	Tempo ao hospital (minutos)	Tempo ao metro (minutos)	Tempo ao aterro (minutos)	Cozinha_equipada	Estacionamento	Varanda	Ar_condicionado	Condominio_fechado
1	1	6,658	1,000	,00	,00	,00	,00	,00	,00	,00	,01	,00	,01	,00	,00
	2	1,428	2,159	,00	,00	,12	,00	,01	,01	,00	,00	,10	,00	,06	,16
	3	,846	2,806	,00	,00	,37	,00	,00	,00	,00	,00	,00	,00	,58	,00
	4	,799	2,886	,00	,00	,23	,00	,01	,00	,00	,00	,00	,03	,15	,58
	5	,669	3,155	,00	,00	,22	,00	,01	,01	,00	,09	,15	,22	,11	,08
	6	,526	3,559	,00	,00	,02	,00	,00	,00	,00	,00	,39	,56	,00	,15
	7	,364	4,274	,00	,05	,02	,03	,05	,12	,00	,26	,04	,01	,06	,00
	8	,338	4,436	,00	,00	,00	,00	,10	,09	,00	,43	,26	,13	,02	,02
	9	,181	6,068	,00	,00	,00	,00	,78	,36	,03	,05	,00	,01	,00	,00
	10	,129	7,184	,03	,07	,02	,05	,03	,38	,15	,10	,03	,01	,00	,01
	11	,041	12,762	,00	,82	,00	,88	,00	,00	,00	,01	,00	,00	,00	,00
	12	,021	17,934	,96	,06	,00	,03	,00	,02	,82	,05	,03	,01	,00	,00

5.4. Complementos

5.4.1. Remoção de *outliers*

É possível usar a amostra de *Studentized Deleted Residuals*, gravada automaticamente pelo SPSS quando executa a regressão linear, para a deteção de *outliers* multivariados (12). Estes resíduos seguem uma distribuição *t-Student* com $565 - 11 - 1 = 553$ graus de liberdade. Para cada um destes resíduos é efetuado um teste de hipóteses: não é um *outlier* (H_0) *vs.* é um *outlier* (H_1). Calculou-se o $p - value$ associado a cada um destes testes, rejeitando-se H_0, de que determinada observação não é um *outlier*, se $p - value \leq .05$. Após remoção dos casos classificados como *outliers* na base de dados, efetuou-se uma nova análise de regressão em tudo semelhante à realizada anteriormente, que conduziu aos resultados que se expõem a seguir. Informação adicional pode ser consultada no Anexo B.

O gráfico de probabilidade Normal dos resíduos construído, Figura 21, fornece boas indicações quanto à Normalidade dos erros.

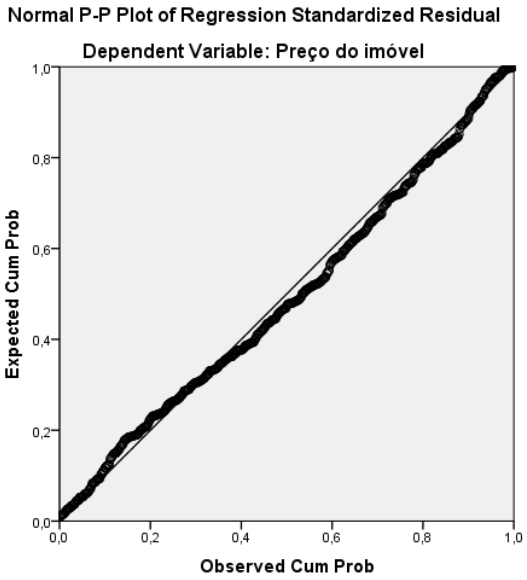

Figura 21- Gráfico de probabilidade Normal.

O gráfico de dispersão de resíduos estandardizados *vs*. valores preditos, Figura 22, apresenta uma mancha de pontos aleatórios, sem qualquer tendência suspeita. Estas características apontam no sentido da independência e variância constante dos erros do modelo.

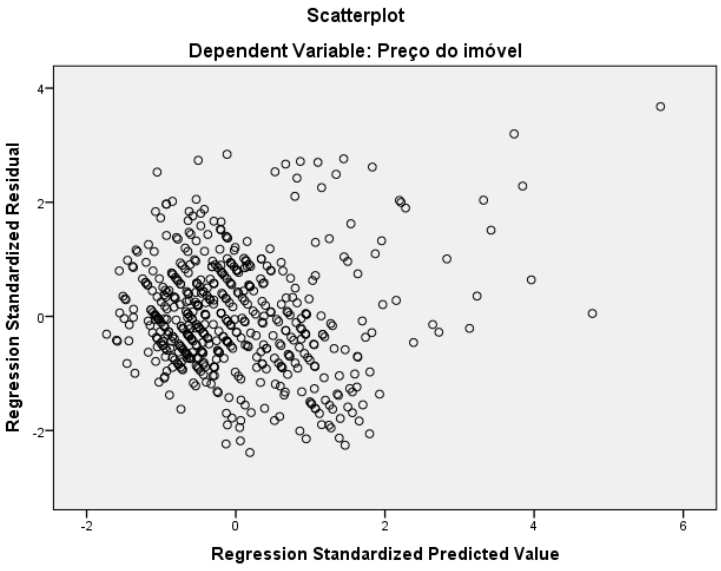

Figura 22- Gráfico dos resíduos estandardizados vs. valores preditos estandardizados.

A estatística de teste de Durbin-Watson é $d \approx 1.823$, pelo que não se rejeita a hipótese nula da ausência de autocorrelação entre os erros do modelo. Os valores de *VIF*, *Tolerance* e *Condition*

Index encontram-se dentro dos valores de referência, não havendo portanto motivos de preocupação com a colinearidade entre as variáveis independentes.

Houve efetivamente uma melhoria do desempenho do modelo de regressão, uma vez que se obteve $\overline{R^2} = 0.720$. Este modelo explica 72% da variabilidade observada no preço de arrendamento. Pela análise do resultado do teste de significância global o modelo revelou-se significativo. Os testes aos coeficientes de regressão permitem concluir que as variáveis **Time_aterro** e **Varanda** não são relevantes considerando $\alpha = 0.05$, mas já o são considerando $\alpha = 0.1$. As conclusões relativas aos sinais dos coeficientes mantêm-se inalteradas. A contribuição relativa de cada variável difere um pouco das obtidas com o Modelo 9. Salienta-se que as variáveis **Estacionamento** e **Time_hospital** se tornaram as terceira e quarta mais relevantes no modelo, ultrapassando a variável **State**.

Finalmente, o modelo ajustado é definido por:

$$\begin{aligned}\widehat{\boldsymbol{Price}} = 196.652 \; + \; 4.5\,\boldsymbol{UsefulArea} \; + \; 98.503\,\boldsymbol{NrWCs} \; + 145.505\,\boldsymbol{State} \\ - 7.286\,\boldsymbol{Time\_subway} + 4.594\,\boldsymbol{Time\_aterro} - 16.713\,\boldsymbol{Time\_hospital} \\ + \; 46.367\,\boldsymbol{Cozinha\_equipada} + 99.022\,\boldsymbol{Estacionamento} \\ + 25.302\,\boldsymbol{Varanda} + 85.303\,\boldsymbol{Ar\_condicionado} \\ + 111.056\,\boldsymbol{Condominio\_fechado}.\end{aligned}$$

Embora esta alternativa esteja aqui a ser apresentada, devemos ter algumas reservas com a sua aplicação. De facto, admitindo que estas observações, identificadas como *outliers*, não são erros na base de dados e que fazem parte do fenómeno em estudo, a sua remoção, obviamente, deixa de refleti-lo. O desacordo com esta estratégia de eliminação de registos foi a motivação necessária que levou à implementação de uma metodologia robusta, que será sucintamente apresentada na secção seguinte e cujos resultados serão posteriormente relacionados com os obtidos através do método dos mínimos quadrados.

5.4.2. Metodologia robusta

Os métodos estatísticos clássicos, como é o caso do método dos mínimos quadrados usado na regressão linear múltipla anterior, fornecem respostas confiáveis quando os dados validam um conjunto de pressupostos. Na realidade estes pressupostos raramente se verificam, pelo que surgiu a necessidade de construir modelos que produzissem resultados estáveis, não só quando os dados são bem comportados, mas também quando existem observações atípicas que se desviam consideravelmente do padrão geral registado. É neste contexto que surge a abordagem robusta.

As estimativas de mínimos quadrados são adversamente influenciadas pela presença de observações influentes, pelo que falham frequentemente no bom ajustamento à maioria dos dados (19). A justificação reside no facto do modelo de regressão linear múltipla (obtido pelo método dos mínimos quadrados) se ajustar a todos os dados, levando a que a existência de observações influentes o afete fortemente, enquanto um modelo de regressão robusta se ajusta somente à maioria dos dados. A ideia principal da regressão robusta reside na atribuição de um peso a cada observação tendo por base a sua influência.

Neste trabalho, escolhe-se o método dos mínimos quadrados reponderados iterativamente, cuja expressão, em notação matricial, é dada por

$$b^{(i+1)} = \left(X'W^{(i)}X\right)^{-1}X'W^{(i)}Y\ ,$$

onde $W^{(i)}$ é uma matriz diagonal $(n \times n)$ de pesos dos resíduos na iteração i. Os pesos atribuídos aos resíduos em cada iteração são calculados através de funções específicas, como, por exemplo, a função biponderada de Tukey, Andrews, Huber ou *bisquare*.

A função peso utilizada foi a *bisquare*. Enquanto no método dos mínimos quadrados é dado um peso de 1 a todas as observações, com esta função cada observação tem um peso atribuído consoante o resíduo correspondente é mais ou menos próximo de zero. Deste modo, quanto mais o resíduo (em valor absoluto) se afastar de zero, menor é o peso atribuído à observação que lhe está associada. Portanto, concluímos que quanto maior o número de casos com peso próximo de um, mais próximas são as estimativas obtidas por regressão robusta das obtidas usando o método dos mínimos quadrados.

Comparando as estimativas dos coeficientes de regressão robusta, Tabela 16, com as do Modelo 9, destaca-se a diferença de magnitude das mesmas, especialmente a referente à constante. Esta situação não é estranha, pois reflete a essência da abordagem robusta anteriormente explicada. Constatamos que os sinais dos coeficientes permanecem inalterados. Por exemplo, uma maior distância a uma estação de metro ou hospital têm um efeito negativo na formação do preço de arrendamento médio de um imóvel, no concelho de Lisboa, tal como já acontecia no Modelo 9.

Tabela 16- Estimativas e significância dos coeficientes obtidos por regressão robusta.

	Estimativas	*p-values*
Constante	280.3129	0.0000
UsefulArea	3.9971	0.0000
NrWCs	73.3589	0.0000
State	129.0190	0.0001
Time_subway	-5.6032	0.0970
Time_aterro	1.4138	0.5899
Time_hospital	-13.4062	0.0019
Cozinha_equipada	44.1055	0.0033
Estacionamento	103.4435	0.0000
Varanda	28.1098	0.0605
Ar_condicionado	62.5546	0.0072
Condominio_fechado	97.6140	0.0060

Analisando a significância dos coeficientes apurados na mesma tabela, voltamos a concluir que as variáveis relacionadas com os atributos físicos continuam a ser relevantes. Relativamente às características de localização, se elevarmos o nível de significância, α, para 0.1, somente deixa de ser relevante a variável **Time_aterro**. Mantendo o nível de significância em 0.05, das características de localização incluídas na regressão, apenas **Time_hospital** é considerada relevante na explicação do comportamento médio do preço dos imóveis.

5.4.3. Método de máxima entropia

Para reforçar os resultados obtidos com o método dos mínimos quadrados, testou-se ainda uma outra metodologia de estimação e inferência denominada máxima entropia generalizada (GME). Este método é apelativo, pois não impõe pressupostos restritivos sobre os dados, como no caso do método dos mínimos quadrados.

Sendo um método pouco conhecido, para conveniência do leitor, o estimador GME é aqui sucintamente descrito (20). Considere-se o modelo linear $Y = X\beta + U$, tal que Y é o vetor $n \times 1$ da variável dependente, X é a matriz $n \times k$ correspondente às variáveis independentes, β é o vetor $k \times 1$ dos coeficientes de regressão e U é o vetor $n \times 1$ de erros. O estimador GME dos coeficientes da regressão, b, é dado por $b = Z\hat{p}$, com $\hat{p} = (\hat{p}_1', \dots, \hat{p}_k')$, solução do problema de otimização assim definido:

$$\underset{p_i, \omega_j : \forall i,j}{\operatorname{argmax}} \left[-\sum_{i=1}^{k} p_i' \, ln(p_i) - \sum_{j=1}^{n} \omega_j' \, ln(\omega_j) \right]$$

$sujeito\ a$:

$$Y = XZp + V\omega$$

$$\mathbf{1}'p_i = \mathbf{1}, \forall i$$

$$\mathbf{1}'\omega_j = \mathbf{1}, \forall j$$

$$p_i > [0], \omega_j > [0], \forall i, j$$

onde $p = (p_1', \dots, p_k')'$. As matrizes Z e V , de dimensões $k \times km$ e $n \times nh$, respetivamente, são matrizes de suporte dos vetores β e U, que se definem como

$$Z = \begin{bmatrix} z_1' & 0 & \dots & 0 \\ 0 & z_2' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & z_k' \end{bmatrix} \quad \text{e} \quad V = \begin{bmatrix} v_1' & 0 & \dots & 0 \\ 0 & v_2' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & v_n' \end{bmatrix},$$

em que $z_i = (z_{i1}, \dots, z_{im})'$ é um vetor $m \times 1$ tal que $z_{i1} \leq z_{i2} \leq \cdots \leq z_{im}$ e $\beta_i \in (z_{i1}, z_{im})$, $\forall i = 1, \dots, k$. De forma análoga, $v_j = \left(v_{j1}, \dots, v_{jh}\right)'$ é um vetor $h \times 1$ tal que $v_{j1} \leq v_{j2} \leq \cdots \leq v_{jh}$ e $U_j \in \left(v_{j1}, v_{jh}\right)$, $\forall j = 1, \dots, n$.

O estimador GME foi implementado considerando diferentes suportes, de diferentes amplitudes, para os parâmetros do modelo, tendo-se verificado uma grande estabilidade das estimativas obtidas perante os diferentes suportes analisados. Os suportes para os erros do modelo foram definidos recorrendo à regra 3σ . Foram considerados $m = 5$ e $h = 5$, isto é, cinco pontos em todos os suportes do modelo. As estimativas dos coeficientes de regressão obtidos com o GME foram semelhantes em magnitude e iguais em sinal às anteriormente conseguidas com o método dos mínimos quadrados, corroborando, portanto, a investigação realizada anteriormente.

Detalhes adicionais sobre o estimador GME podem ser consultados na obra de Golan, Judge e Miller (21).[2]

[2] Desenvolvimentos recentes nesta área podem ser consultados, por exemplo, no Info-Metrics Institute (http://www.american.edu/cas/economics/info-metrics/).

6. Conclusões e considerações finais

Com o objetivo de construir um modelo estatístico que permita identificar os atributos determinantes do preço de arrendamento de um imóvel destinado a habitação, no concelho de Lisboa, iniciou-se o estudo com uma análise exploratória, a qual consistiu na execução de sucessivas regressões com os seguintes métodos de seleção de variáveis: *stepwise*, *backward*, *forward* e entrada forçada. Esta fase levou à seleção de 11 variáveis respeitantes a atributos físicos e de localização. Todas estão presentes no modelo de regressão linear múltipla final e que foi alvo de uma análise mais aprofundada. Durante todo o processo de análise foi considerado um nível de significância $\alpha = 0.05$.

Para a deteção de *outliers* e casos influentes foram examinados os resíduos estandardizados, os valores da estatística da distância de Cook e ainda os *DFBetas*. Não se revelou óbvia a eliminação de registos da base de dados.

Posteriormente foram analisados os pressupostos do modelo de regressão linear múltipla, nomeadamente o da Normalidade, homogeneidade e independência dos resíduos, bem como o da ausência de colinearidade entre as variáveis independentes. Os dois primeiros foram tratados graficamente e com recurso aos testes de ajustamento de Kolmogorov-Smirnov com correção de Lilliefors e de White, respetivamente. Em ambos os casos, *os* $p - values$ associados aos testes levaram à rejeição da hipótese nula, portanto, da Normalidade e da homocedasticidade dos erros. No entanto, o facto do pressuposto da Normalidade não se verificar não anula a inferência realizada, pois lidamos com uma amostra de dimensão elevada (11) . Por outro lado, a ausência de homocedasticidade não invalida a análise, apenas a enfraquece (17). O pressuposto da ausência de autocorrelação foi validado com a estatística de Durbin-Watson. Para diagnosticar eventuais dificuldades com colinearidade foram analisados os valores de *VIF*, *Tolerance* e, ainda, de *Condition Index*. Neste âmbito, para a amostra e modelo em causa, a conclusão é a da não existência de colinearidade entre as variáveis independentes.

6.1. Principais conclusões

A regressão linear múltipla levou à identificação de 11 variáveis consideradas relevantes do comportamento médio de **Price**, o preço de arrendamento. O modelo final considerado neste trabalho é

$$\widehat{\boldsymbol{Price}} = 93.733 + 5.132\,\boldsymbol{UsefulArea} + 100.224\,\boldsymbol{NrWCs} + 213.707\,\boldsymbol{State} - 10.129\,\boldsymbol{Time\_subway} + 8.401\,\boldsymbol{Time\_aterro} - 13.689\,\boldsymbol{Time\_hospital} + 56.197\,\boldsymbol{Cozinha\_equipada} + 98.401\,\boldsymbol{Estacionamento} + 50.454\,\boldsymbol{Varanda} + 137.894\,\boldsymbol{Ar\_condicionado} + 146.790\,\boldsymbol{Condominio\_fechado}.$$

O modelo de regressão acima mencionado explica cerca de 65,6% da variação observada na variável dependente **Price**, em torno da sua média. Avaliando os valores absolutos dos coeficientes de regressão estandardizados (Tabela 14), concluímos que os atributos físicos área útil, número de casas de banho e estado do imóvel são os que apresentam maior contribuição relativa para explicar o comportamento esperado do preço de arrendamento. Seguem-se as variáveis selecionadas relativas à presença ou ausência de características específicas (exceto, a existência de varanda) e, finalmente, as variáveis respeitantes à localização do imóvel.

6.2. Considerações finais

Com o trabalho desenvolvido pretendeu-se investigar a importância de atributos físicos e de localização de um imóvel na determinação do seu preço de arrendamento, no concelho de Lisboa. A grande dificuldade foi, indubitavelmente, a base de dados. A preparação dos dados para que posteriormente pudessem ser trabalhados implicou um esforço acrescido no domínio de ferramentas como o Excel e o SPSS Statistics. De facto, a qualidade dos dados foi, durante todo o estudo, um aspeto que suscitou dúvidas. A sua origem, no portal imobiliário Imovirtual, é o seu ponto mais fraco, pois estamos perante um procedimento pouco controlado. Neste âmbito, salienta-se a falta de uniformização no preenchimento da ficha de um imóvel quando este é inserido. A base de dados revelou-se, portanto, confusa e pouco fidedigna.

No que diz respeito à recolha de informação sobre as características de localização de um imóvel, destaca-se como principal obstáculo a falta de conhecimento, de uma forma exaustiva, de todos os serviços, amenidades e locais de reconhecida influência na tomada de decisão da escolha de uma habitação para arrendamento, presentes na sua área geográfica. Por outro lado, a informação relativa aos tempos e distâncias são pouco exatas, uma vez que foram obtidas através do código postal dos imóveis e não a partir dos seus endereços completos.

Tal como era expectável, revelaram-se importantes para explicar o comportamento médio dos preços de arrendamento, variáveis relacionadas com atributos físicos dos imóveis e variáveis relacionadas com a localização. No entanto, estas últimas não demonstraram ter o impacto

inicialmente esperado, quiçá, devido à fragilidade da informação já mencionada. Como seria de esperar, as variáveis respeitantes à área útil, ao número de casas de banho e à condição do imóvel são as que mostraram ter maior contribuição relativa no comportamento do preço médio de arrendamento. Apesar da reduzida influência das variáveis de localização, não deixa de ser interessante constatar que as variáveis selecionadas desta componente sejam **Time_aterro**, **Time_hospital** e **Time_subway**. Acresce ainda referir que o número de regressores incluídos no modelo final é muito inferior ao número de variáveis independentes inicial. A redução deste número foi superior a 50% e deveu-se, sobretudo, à aplicação de métodos de seleção de variáveis, não obstante alguma sensibilidade pessoal sobre o tema. Neste contexto, sugere-se, para trabalho futuro, a colaboração de um profissional que domine o mercado imobiliário, a fim de se conseguir chegar a uma possível melhor combinação de atributos para explicar a variação dos preços de arrendamento.

Propõe-se, igualmente, para trabalho futuro, a inclusão de variáveis como o Imposto Municipal sobre Imóveis (IMI), por ser um indicador usual de avaliação do valor patrimonial de um imóvel. Se observarmos que o cálculo do valor de IMI de um imóvel envolve, entre outros, coeficientes de localização, vetustez, qualidade e conforto, é fácil perceber a riqueza deste tipo de índice num estudo como o efetuado.

Como foi explicado ao longo do trabalho, a variável concernente à idade não foi tida em conta por apresentar um número elevado de valores omissos. No entanto, seria interessante perceber em que medida ela influencia o preço de arrendamento, pois poderia, eventualmente, traduzir de forma mais clara a importância da condição do imóvel. A variável **State** é claramente insuficiente para esse efeito.

Por outro lado, uma vez que o estudo incidiu sobre o concelho de Lisboa, sendo o trânsito um problema incontornável desta zona geográfica, seria curioso perceber quais as conclusões obtidas com a inclusão de informação respeitante ao tráfego. Ou seja, tendo em conta os fracos resultados obtidos pelas variáveis de localização experimentadas, tentar agregar a informação das distâncias com tempos reais de deslocação e custos da mesma, talvez fosse uma mais-valia para estudos futuros.

Referências bibliográficas

1. Batista P. O data mining na identificação de atributos valorativos da habitação [Internet] [dissertation on the Internet]. [Aveiro, Portugal]: Universidade de Aveiro, Secção Autónoma de Ciências Sociais Jurídicas e Políticas; 2010 [cited 2015 Jan 15]. Available from: http://hdl.handle.net/10773/3640

2. Costa JS. Compêndio de Economia Regional. Coimbra: Associação Portuguesa para o Desenvolvimento Regional; 2002. 851 p.

3. Bastos IM. Dinâmica de Preços no Mercado da Habitação : Análise de Clusters Aplicada à Cidade do Porto [Internet] [dissertation on the Internet]. [Porto, Portugal]: Universidade do Porto, Faculdade de Economia; 2013 [cited 2015 Jan 15]. Available from: http://repositorio-aberto.up.pt/handle/10216/70547

4. Couto PM. Avaliação Patrimonial de Imóveis para Habitação [Internet] [master's thesis on the Internet]. [Porto, Portugal]: Universidade do Porto, Faculdade de Engenharia; 1999 [cited 2015 Jan 15]. Available from: http://repositorio-aberto.up.pt/handle/10216/12152

5. Tarré AFM do V. Análise de valores de avaliação de apartamentos no âmbito do Crédito a Habitação, para duas zonas distintas do concelho de Lisboa- recurso a Modelos Hedónicos [Internet] [dissertation on the internet]. [Lisboa, Portugal]: Universidade Técnica de Lisboa, Instituto Superior de Economia e Gestão; 2009 [cited 2015 May 10]. Available from: https://www.repository.utl.pt/handle/10400.5/1296

6. Gonzalez MAS, Formoso CT. Estimativa de modelos de preços hedônicos para locação residencial em Porto Alegre. Produção. 1995;5(1):65–77.

7. Guedes TB. Polinómios fracionários na modelação do preço de imóveis [Internet] [dissertation on the Internet]. [Aveiro, Portugal]: Universidade de Aveiro, Departamento de Matemática; 2011 [cited 2015 Jan 15]. Available from: http://ria.ua.pt/handle/10773/10493

8. Neto F da S. Aplicação de um modelo hedónico de avaliação a edifícios habitacionais no concelho de Gaia [Internet] [dissertation on the internet]. [Lisboa, Portugal]: Universidade Técnica de Lisboa, Instituto Superior de Economia e Gestão; 2008 [cited 2015 May 5]. Available from: https://www.repository.utl.pt/handle/10400.5/532

9. Tavares F, Moreira A, Pereira E. Avaliação imobiliária:Dois casos da importância das vistas como externalidades. Rev Port e Bras Gestão. 2012;2–13.

10. Tavares FA, Moreira AC, Pereira ET. Avaliação Imobiliária sob a Perspectiva das Externalidades: uma revisão da literatura. Universo Contábil. 2010;6(3):96–113.

11. Murteira B, Ribeiro CS, Silva JA e, Pimenta C. Introdução à estatística. Lisboa: Escolar Editora; 2010. 696 p.

12. Marôco J. Análise Estatística com o SPSS Statistics. 6th ed. Pêro Pinheiro: ReportNumber; 2014. 990 p.

13. Hall A, Neves C, Pereira A. Grande Maratona de Estatística no SPSS. Lisboa: Escolar Editora; 2011. 360 p.

14. Guimarães RC, Cabral JA. Estatística. 2nd ed. Madrid: McGraw-Hill; 2007. 455 p.

15. Field A. Discovering Statistics using SPSS. 3rd ed. London: Sage Publications; 2009. 822 p.

16. Stevens JP. Applied multivariate statistics for the social sciences. 5th ed. New York: Routledge; 2009. 651 p.

17. Tabachnick BG, Fidell LS. Using Multivariate Statistics. 4th ed. New York: Harper Collins; 2001. 966 p.

18. White H. A heteroscedasticity- consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrica. 1980;48(4):817–38.

19. Maronna RA, Martin RD, Yohai VJ. Robust Statistics Theory and Methods. Chichester: John Wiley; 2006. 403 p.

20. Mittelhammer R, Cardell NS, Marsh TL. The data-constrained generalized maximum entropy estimator of the GLM: Asymptotic theory and inference. Entropy. 2013;15(5):1756–75.

21. Golan A, Judge GG, Miller D. Maximum entropy econometrics: robust estimation with limited data. Chichester: Wiley; 1996. 324 p.

Anexos

A. *Outputs* dos modelos de regressão linear múltipla

- MODELO 1

Tabela 17- Estatísticas do Modelo 1 estimado.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,797[a]	,636	,628	204,4594	,636	82,922	12	570	,000	1,899

Tabela 18- Estatísticas do Modelo 1 estimado (continuação).

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	182,632	70,726		2,582	,010	43,717	321,548					
	Número de quartos	20,715	11,624	,079	1,782	,075	-2,117	43,547	,589	,074	,045	,322	3,110
	Área útil	4,668	,465	,521	10,029	,000	3,754	5,582	,719	,387	,253	,237	4,219
	Número de WCs	92,711	21,555	,163	4,301	,000	50,374	135,049	,627	,177	,109	,442	2,261
	Soma das caracteristicas	36,706	3,651	,275	10,055	,000	29,536	43,876	,366	,388	,254	,855	1,170
	Tempo ao café (minutos)	-16,436	9,212	-,071	-1,784	,075	-34,530	1,657	-,033	-,075	-,045	,408	2,453
	Tempo ao hospital (minutos)	-14,989	7,624	-,079	-1,966	,050	-29,963	-,016	-,070	-,082	-,050	,395	2,535

Tempo à refinaria (minutos)	,375	3,878	,003	,097	,923	-7,242	7,992	,004	,004	,002	,629	1,590
Tempo ao restaurante (minutos)	13,627	6,846	,060	1,991	,047	,181	27,073	,099	,083	,050	,714	1,401
Tempo à escola (minutos)	-20,023	9,998	-,060	-2,003	,046	-39,660	-,387	-,012	-,084	-,051	,702	1,425
Tempo ao metro (minutos)	-12,544	5,135	-,084	-2,443	,015	-22,629	-2,459	-,120	-,102	-,062	,535	1,869
Tempo ao aterro (minutos)	11,430	3,953	,096	2,891	,004	3,665	19,195	-,184	,120	,073	,580	1,725
Número de pontos de interesse(raio 1Km)	-12,696	4,989	-,098	-2,545	,011	-22,494	-2,898	-,045	-,106	-,064	,428	2,337

- MODELO 2

Tabela 19- Estatísticas do Modelo 2 estimado.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,796[a]	,633	,630	214,9549	,633	210,095	5	608	,000	1,813

Tabela 20- Estatísticas do Modelo 2 estimado (continuação).

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	40,591	50,258		,808	,420	-58,110	139,291					
	Área útil	5,135	,349	,546	14,732	,000	4,451	5,820	,720	,513	,362	,440	2,274
	Número de WCs	105,551	21,802	,180	4,841	,000	62,734	148,368	,643	,193	,119	,437	2,290
	Soma das caracteristicas	41,866	3,461	,312	12,098	,000	35,070	48,663	,426	,440	,297	,908	1,102
	Tempo ao aterro (minutos)	5,187	3,267	,041	1,588	,113	-1,228	11,602	-,189	,064	,039	,894	1,118
	Número de pontos de interesse(raio 1Km)	,668	3,294	,005	,203	,839	-5,802	7,137	-,061	,008	,005	,982	1,018

- MODELO 3

Tabela 21- Estatísticas do Modelo 3 estimado.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,810[a]	,656	,649	205,1922	,656	94,833	11	547	,000	1,901

Tabela 22- Estatísticas do Modelo 3 estimado (continuação).

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	231,517	63,687		3,635	,000	106,415	356,619					
	Número de quartos	21,585	11,999	,080	1,799	,073	-1,985	45,155	,576	,077	,045	,315	3,173
	Área útil	4,500	,487	,485	9,235	,000	3,543	5,457	,716	,367	,232	,228	4,392
	Estado do imóvel	225,963	42,873	,140	5,271	,000	141,747	310,179	,316	,220	,132	,891	1,122
	Número de WCs	98,217	21,981	,169	4,468	,000	55,040	141,394	,636	,188	,112	,438	2,283
	Soma das caracteristicas	34,613	3,826	,252	9,048	,000	27,099	42,128	,396	,361	,227	,811	1,233
	Tempo ao café (minutos)	-17,062	9,039	-,072	-1,888	,060	-34,817	,693	-,044	-,080	-,047	,434	2,305
	Tempo ao hospital (minutos)	-10,829	7,477	-,055	-1,448	,148	-25,516	3,858	-,072	-,062	-,036	,434	2,303
	Tempo à escola (minutos)	-19,690	10,087	-,058	-1,952	,051	-39,504	,125	-,027	-,083	-,049	,723	1,383
	Tempo ao metro (minutos)	-14,718	4,590	-,096	-3,206	,001	-23,734	-5,701	-,127	-,136	-,080	,697	1,434
	Tempo ao aterro (minutos)	9,913	3,682	,080	2,692	,007	2,680	17,146	-,190	,114	,068	,711	1,406
	Número de pontos de interesse(raio 1Km)	-14,905	5,098	-,112	-2,924	,004	-24,919	-4,890	-,044	-,124	-,073	,430	2,325

- MODELO 4

Tabela 23- Estatísticas do Modelo 4 estimado.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,806[a]	,650	,644	206,1330	,650	113,953	9	553	,000	1,890

Tabela 24- Estatísticas do Modelo 4 estimado (continuação).

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	141,148	57,084		2,473	,014	29,020	253,277					
	Número de quartos	20,860	12,034	,078	1,733	,084	-2,778	44,498	,575	,074	,044	,316	3,163
	Área útil	4,478	,488	,483	9,169	,000	3,519	5,438	,716	,363	,231	,228	4,378
	Estado do imóvel	227,351	43,058	,141	5,280	,000	142,773	311,928	,316	,219	,133	,892	1,122
	Número de WCs	95,941	22,054	,165	4,350	,000	52,621	139,261	,637	,182	,109	,438	2,285
	Soma das caracteristicas	36,322	3,801	,264	9,556	,000	28,856	43,789	,396	,376	,241	,827	1,209
	Tempo à escola (minutos)	-17,095	10,034	-,050	-1,704	,089	-36,805	2,615	-,027	-,072	-,043	,737	1,357
	Tempo ao metro (minutos)	-18,408	4,306	-,123	-4,275	,000	-26,865	-9,951	-,133	-,179	-,108	,764	1,309
	Tempo ao aterro (minutos)	9,438	3,592	,077	2,628	,009	2,383	16,493	-,194	,111	,066	,740	1,350
	Número de pontos de interesse(raio 1Km)	-5,448	4,197	-,041	-1,298	,195	-13,692	2,795	-,041	-,055	-,033	,638	1,566

- MODELO 5

Tabela 25- Estatísticas do Modelo 5 estimado.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,811[a]	,657	,650	205,0197	,657	87,237	12	546	,000	1,901

Tabela 26- Estatísticas do Modelo 5 estimado (continuação).

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	335,178	98,200		3,413	,001	142,281	528,075					
	Número de quartos	20,788	12,003	,077	1,732	,084	-2,790	44,365	,576	,074	,043	,314	3,181
	Área útil	4,547	,488	,490	9,317	,000	3,588	5,506	,716	,370	,233	,227	4,413
	Estado do imóvel	226,876	42,842	,141	5,296	,000	142,721	311,032	,316	,221	,133	,891	1,122
	Número de WCs	97,229	21,974	,168	4,425	,000	54,065	140,393	,636	,186	,111	,437	2,286
	Soma das caracteristicas	34,621	3,822	,252	9,058	,000	27,113	42,130	,396	,361	,227	,811	1,233
	Tipo de imóvel	-95,148	68,651	-,036	-1,386	,166	-230,000	39,704	,042	-,059	-,035	,908	1,101
	Tempo ao café (minutos)	-17,234	9,032	-,073	-1,908	,057	-34,976	,508	-,044	-,081	-,048	,434	2,305
	Tempo ao hospital (minutos)	-11,794	7,503	-,060	-1,572	,117	-26,533	2,944	-,072	-,067	-,039	,431	2,323
	Tempo à escola (minutos)	-19,975	10,081	-,058	-1,981	,048	-39,777	-,173	-,027	-,084	-,050	,723	1,383

Tempo ao metro (minutos)	-15,977	4,675	-,105	-3,417	,001	-25,162	-6,793	-,127	-,145	-,086	,671	1,490
Tempo ao aterro (minutos)	9,878	3,679	,080	2,685	,007	2,651	17,105	-,190	,114	,067	,711	1,406
Número de pontos de interesse(raio 1Km)	-15,488	5,111	-,116	-3,030	,003	-25,528	-5,448	-,044	-,129	-,076	,427	2,340

- MODELO 6

Tabela 27- Estatísticas do Modelo 6 estimado.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,806[a]	,650	,644	206,1330	,650	113,953	9	553	,000	1,890

Tabela 28-Estatísticas do Modelo 6 estimado (continuação).

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	141,148	57,084		2,473	,014	29,020	253,277					
	Número de quartos	20,860	12,034	,078	1,733	,084	-2,778	44,498	,575	,074	,044	,316	3,163
	Área útil	4,478	,488	,483	9,169	,000	3,519	5,438	,716	,363	,231	,228	4,378
	Estado do imóvel	227,351	43,058	,141	5,280	,000	142,773	311,928	,316	,219	,133	,892	1,122
	Número de WCs	95,941	22,054	,165	4,350	,000	52,621	139,261	,637	,182	,109	,438	2,285

Soma das caracteristicas	36,322	3,801	,264	9,556	,000	28,856	43,789	,396	,376	,241	,827	1,209
Tempo à escola (minutos)	-17,095	10,034	-,050	-1,704	,089	-36,805	2,615	-,027	-,072	-,043	,737	1,357
Tempo ao metro (minutos)	-18,408	4,306	-,123	-4,275	,000	-26,865	-9,951	-,133	-,179	-,108	,764	1,309
Tempo ao aterro (minutos)	9,438	3,592	,077	2,628	,009	2,383	16,493	-,194	,111	,066	,740	1,350
Número de pontos de interesse(raio 1Km)	-5,448	4,197	-,041	-1,298	,195	-13,692	2,795	-,041	-,055	-,033	,638	1,566

- MODELO 7

Tabela 29- Estatísticas do Modelo 7 estimado.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,812[a]	,659	,653	203,3134

Tabela 30- Estatísticas do Modelo 7 estimado (continuação).

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	83,489	48,089		1,736	,083
	Área útil	5,083	,351	,548	14,474	,000
	Estado do imóvel	217,817	42,552	,135	5,119	,000
	Número de WCs	98,778	21,833	,170	4,524	,000
	Tempo ao metro (minutos)	-14,923	3,943	-,100	-3,785	,000
	Tempo ao aterro (minutos)	8,098	3,424	,066	2,365	,018
	Cozinha_equipada	57,498	19,502	,083	2,948	,003
	Estacionamento	102,902	26,287	,115	3,915	,000

Varanda	48,174	19,511	,066	2,469	,014
Ar_condicionado	141,411	30,260	,122	4,673	,000
Condominio_fechado	143,343	46,208	,082	3,102	,002

- MODELO 8

Tabela 31- Estatísticas do Modelo 8 estimado.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,815[a]	,665	,658	201,9433

Tabela 32- Estatísticas do Modelo 8 estimado (continuação).

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	143,019	54,683		2,615	,009
	Área útil	5,140	,349	,554	14,710	,000
	Estado do imóvel	215,688	42,311	,133	5,098	,000
	Número de WCs	98,994	21,703	,171	4,561	,000
	Tempo ao metro (minutos)	-11,277	4,419	-,076	-2,552	,011
	Tempo ao aterro (minutos)	10,247	3,543	,084	2,892	,004
	Cozinha_equipada	56,079	19,378	,081	2,894	,004
	Estacionamento	99,340	26,179	,111	3,795	,000
	Varanda	49,616	19,407	,068	2,557	,011
	Ar_condicionado	139,322	30,099	,120	4,629	,000
	Condominio_fechado	140,374	46,046	,080	3,049	,002
	Tempo ao hospital (minutos)	-19,523	6,385	-,100	-3,058	,002
	Número de pontos de interesse(raio 1Km)	-7,667	4,090	-,058	-1,875	,061

B. *Outputs* dos modelos de regressão linear múltipla (sem *outliers*)

Tabela 33- Estatísticas do Modelo 9 (sem outliers).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,852[a]	,726	,720	137,2813	1,823

Tabela 34- ANOVA do Modelo 9 (sem outliers).

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	26108763,118	11	2373523,920	125,942	,000[b]
	Residual	9856535,200	523	18846,148		
	Total	35965298,318	534			

Tabela 35- Coeficientes de regressão estimados do Modelo 9 (sem outliers).

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	196,652	34,300		5,733	,000		
	Área útil	4,500	,271	,589	16,594	,000	,416	2,406
	Número de WCs	98,503	16,094	,214	6,120	,000	,427	2,342
	Estado do imóvel	145,505	31,985	,109	4,549	,000	,913	1,095
	Tempo ao metro (minutos)	-7,286	3,074	-,065	-2,370	,018	,693	1,443
	Tempo ao aterro (minutos)	4,594	2,391	,050	1,921	,055	,759	1,318
	Tempo ao hospital (minutos)	-16,713	3,904	-,113	-4,281	,000	,748	1,338
	Cozinha_equipada	46,367	13,440	,089	3,450	,001	,782	1,279
	Estacionamento	99,022	18,732	,143	5,286	,000	,711	1,406
	Varanda	25,302	13,593	,046	1,861	,063	,861	1,161
	Ar_condicionado	85,303	22,523	,089	3,787	,000	,939	1,064
	Condominio_fechado	111,056	34,741	,077	3,197	,001	,898	1,114

Tabela 36- Estatísticas de diagnóstico de colinearidade do Modelo 9 (sem outliers).

Collinearity Diagnostics

Model		Eigenvalue	Condition Index	Variance Proportions											
				(Constant)	Área útil	Número de WCs	Estado do imóvel	Tempo ao metro	Tempo ao aterro	Tempo ao hospital	Cozinha_equipada	Estacionamento	Varanda	Ar_condicionado	Condominio_fechado
1	1	6,580	1,000	,00	,00	,00	,00	,00	,00	,00	,01	,00	,01	,00	,00
	2	1,318	2,234	,00	,00	,00	,08	,01	,00	,01	,01	,14	,00	,06	,18
	3	,977	2,596	,00	,00	,00	,57	,00	,00	,00	,00	,00	,01	,12	,18
	4	,865	2,758	,00	,00	,00	,00	,00	,00	,00	,00	,00	,01	,63	,31
	5	,668	3,138	,00	,00	,00	,25	,01	,00	,01	,08	,06	,32	,12	,09
	6	,520	3,557	,00	,00	,00	,06	,00	,00	,00	,01	,44	,46	,01	,21
	7	,360	4,275	,00	,04	,03	,02	,11	,00	,07	,27	,04	,01	,04	,00
	8	,343	4,379	,00	,00	,00	,01	,09	,00	,10	,45	,24	,16	,02	,02
	9	,179	6,056	,00	,00	,00	,00	,36	,03	,77	,03	,00	,01	,00	,00
	10	,132	7,065	,03	,05	,05	,01	,38	,14	,03	,08	,03	,01	,00	,01
	11	,038	13,220	,00	,79	,90	,00	,00	,00	,00	,01	,00	,00	,00	,00
	12	,020	18,266	,96	,11	,01	,00	,03	,82	,00	,06	,03	,01	,00	,00